Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Estudi de mètodes de selecció automàtica de variables aplicats a la predicció de resultats acadèmics

MEMÒRIA

Autor: Carles Llonch Ciruelo

Director: Luis José Talavera Méndez

Convocatòria: Juny de 2019

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

RESUM

L'objectiu d'aquest treball és estudiar i aplicar mètodes de selecció automàtica de variables utilitzats per la creació de models de predicció de resultats acadèmics d'estudiants de l'ETSEIB. En concret, s'intenta predir l'aprovat o suspès de les assignatures del tercer quadrimestre del grau en enginyeria en tecnologies industrials, donat que són les primeres assignatures a cursar un cop superada la fase inicial.

Per dur a terme aquesta tasca s'aplica de forma rigorosa la metodologia CRISP-DM per al desenvolupament de projectes de mineria de dades. Tot plegat comptant amb el suport de les eines proporcionades per les llibreries *Pandas* i *SciKit-learn* del llenguatge *Python*.

Les dades d'estudi són proporcionades per la pròpia escola, les quals contenen informació dels resultats acadèmics de cada estudiant així com algunes altres dades personals. Amb aquestes mateixes es validen els models de predicció de regressió logística i arbre de decisions, els quals es compara el rendiment obtingut en la seva resposta.

Sumari

1. Prefaci

Origen del projecte

El projecte escollit sorgeix de la borsa de Treballs de Fi d'Estudis que facilita l'ETSEIB. Del gran nombre de possibilitats que s'oferien a la plataforma, es buscava un treball focalitzat en l'ús d'eines de coneixement a nivell matemàtic.

Així doncs, el departament de Ciències de la Computació va resultar ser el més atractiu, ja permetia desenvolupar recursos matemàtics i estadístics mitjançant l'ús experimental de la programació. L'estudi i anàlisi de dades va ser un dels temes més interessants, ja que avui en dia es considera una eina primordial en qualsevol sector industrial o informàtic.

Motivació

Els motius que van decantar la balança per escollir aquest tema com a Treball de Fi de Grau van ser la finalitat i el desenvolupament d'aquest. Per una banda, va resultar motivador el fet de poder realitzar un estudi amb les dades d'estudiants de la pròpia escola, així com intentar oferir una visió global del seu rendiment acadèmic, quelcom que tot estudiant ha intentat avaluar algun dia.

Per altra banda, el repte de realitzar l'estudi mitjançant el llenguatge de programació *Python* resultava molt atractiu, ja que es podria recuperar allò après els primers anys de grau i profunditzar en un llenguatge molt utilitzat en el món de la programació.

2. Introducció

2.1. Objectius del projecte

L'objectiu principal d'aquest projecte és formular i analitzar models de predicció de rendiment acadèmic pels estudiants del grau d'enginyeria industrial a l'ETSEIB mitjançant tècniques de mineria de dades. En concret, s'estudiaran formes de selecció automàtica de variables rellevants per poder predir l'aprovat o suspès de les assignatures del tercer quadrimestre, és a dir, el primer quadrimestre a cursar un cop superada la fase inicial. A partir dels resultats obtinguts, es farà una avaluació del model de predicció creat i s'analitzarà la resposta obtinguda en funció de les variables seleccionades, amb l'objectiu de poder utilitzar el model en un futur amb casos reals.

Per fer-ho possible, es disposa de dades com els resultats acadèmics de tots els estudiants i estudiantes matriculats a partir de l'any 2010, any de transició al actual pla Bolonya. També es disposa d'algunes altres dades personals dels individus sota estudi, com ara el codi postal, el sexe o bé la nota d'accés de les PAU.

Com a objectiu addicional es pretén adquirir coneixements sobre l'ús de metodologies de mineria de dades, així com aprendre diverses eines de programació que ofereix el llenguatge *Python*. Això permetrà dur a terme un anàlisi de dades més ordenat, hàbil i fiable en un futur professional en el sector de l'enginyeria.

2.2. Abast del projecte

El projecte es realitzarà adaptant la metodologia CRISP-DM a les dimensions i objectius del treball, ja que és un dels mètodes més usats a la mineria de dades. Pel que fa als models de predicció s'utilitzaran la regressió logística i els arbres de decisions, considerats els models més aptes i adequats pels resultats que es volen predir. Posteriorment es realitzarà una comparativa entre ells per veure quin obté millor rendiment i robustesa. D'un ampli ventall de mitjans de selecció de variables, s'utilitzarà un mètode de filtratge representatiu proveït per la llibreria *scikit-learn* de *Python*, donat que permet un marge de joc amb els models de predicció i és de gran compatibilitat amb el llenguatge de programació usat.

Com ja s'ha comentat anteriorment, la base de treball en que es du a terme el projecte és la informació acadèmica i personal dels estudiants i estudiantes de la pròpia escola. Aquest fet indica que no es tracta d'un cas a nivell teòric sinó que se'n poden extreure conclusions útils per posar en pràctica amb els resultats obtinguts. Això pot incloure propostes de canvis i solucions a alguns trets concrets del programa acadèmic de l'escola, afectant així a la trajectòria de centenars d'alumnes que es matriculen cada any. Heus aquí el potencial i transcendència del treball a realitzar.

Cal remarcar que pel bé de la protecció de dades de l'alumnat, les dades proveïdes per l'escola no inclouen cap número d'identitat com DNI o NIF que permeti reconèixer a cap individu objecte d'estudi.

3. Mineria de dades

3.1. Definició

La mineria de dades (en anglès *Data mining*), és un conjunt de tècniques i tecnologies que permeten explorar dades, de manera automàtica o semi-automàtica, amb l'objectiu de trobar patrons repetitius, tendències o regles que expliquin el comportament de les dades en un context determinat. Aquestes tècniques es recolzen sobre els camps de l'estadística i les ciències de la computació, utilitzant mètodes com la intel·ligència artificial, l'aprenentatge autònom (*Machine Learning*) o les xarxes neuronals.

La tasca principal de la mineria de dades és extreure patrons interessants, fins ara desconeguts, per processar-los en informació útil. Aquest nou increment de coneixement podrà ser utilitzat per elaborar sistemes de presa de decisions o per utilitzar els propis models de predicció creats.

3.1.1. Ús de la mineria de dades

La mineria de dades és utilitzada en múltiples accions del nostre dia a dia, des de casos senzills i rutinaris fins a estudis complexes utilitzats a nivell empresarial o estatal.

Un clar exemple és l'ús del codi de barres en articles d'una botiga. Aquests establiments poden processar ràpidament les compres realitzades i utilitzar els sistemes informàtics per determinar els preus dels productes comprats. Addicionalment el sistema operatiu permet mantenir un registre de l'inventari en tot instant, oferint així la possibilitat de contactar amb el proveïdor d'un producte quan aquest s'estigui esgotant. Fins i tot permet dur a terme un balanç econòmic exacte dels guanys i despeses de qualsevol període de temps desitjat. **[1]**

La mineria de dades s'utilitza sobretot en el món de l'empresa, ja que permet analitzar detingudament el comportament del consumidor i preveure quines seran les pròximes accions d'aquest. Per exemple, enregistrant i analitzant les dades adequades de diversos punts de venda d'un producte es podria arribar a predir el comportament del consumidor segons diferents variables, com ara la demografia del lloc, la competència que l'envolta, la visibilitat del producte de cara al client o fins i tot l'estatus social del barri on es troba el punt de venda. Tota aquesta informació pot ser aprofitada per l'empresa per desenvolupar

promocions i nous productes que atreguin l'atenció del consumidor.

Un altre exemple d'un àmbit on la mineria de dades agafa protagonisme són les campanyes electorals. Els diferents partits polítics han adoptat diverses estratègies de captació de vots. Per exemple, s'ha arribat a la conclusió que no té cap sentit esforçar-se en atreure a dos tipus de votants concrets: per un lloc, els que estan molt allunyats políticament i és molt improbable que canviïn d'idea (vots donats per perduts), i per altre banda, els molt pròxims ideològicament (vots donats per guanyats). Així doncs l'estratègia recau en localitzar els votants més indecisos per atreure'ls cap al partit desitjat. Aquesta tasca és més fàcil de predir gràcies a les tecnologies actuals, on tots els actes que fem queden enregistrats en una base de dades, com per exemple el canal de televisió que veiem o el diari digital que llegim.

En referència a la ciència i la tecnologia, la mineria de dades pren una important rellevància tant en el món de les ciències de la salut com en el de l'enginyeria. Ambdós sectors utilitzen el *Data Mining* per analitzar múltiples dades corresponents a diferents variables i crear models de predicció el màxim acurats possible. D'aquesta manera es poden predir càncers segons diferents variables d'informació del pacient, o també analitzar quines són les variables que més afecten a la qualitat d'una peça fabricada amb una màquina industrial.

Focalitzant-nos en el tema d'aquest treball d'estudi, també s'utilitza la mineria de dades per analitzar paràmetres del sistema educatiu, com el cas de la taxa d'abandonament universitari a l'Argentina, que ha arribat a valors de fins el 80% dels estudiants matriculats en algunes universitats. Davant aquestes dades alarmants, una facultat de la *Universidad Tecnológica Nacional* va realitzar un estudi per obtenir patrons indicadors d'abandonament, basant-se en dades acadèmiques i socioeconòmiques dels estudiants del grau en enginyeria en tecnologies d'informació que ofereix la pròpia facultat. A partir de diferents variables dels estudiants com el nombre d'assignatures aprovades per curs, la situació laboral familiar o l'edat d'inici d'estudis, es va poder dur a terme un projecte *Data Mining* per extreure'n conclusions. Entre els resultats més interessants es va poder detectar que l'assistència a classe variava en funció de l'any cursat i la localitat de procedència. També es va poder observar que la quantitat d'assignatures aprovades era diferent segons el nivell d'estudis dels pares de l'estudiant. **[2]**

No obstant, tota la informació extreta de la mineria de dades no es processa de manera automàtica, doncs ha d'haver-hi algú responsable de configurar i tractar les dades objectes

d'estudi. Aquesta labor es basa en procediments estructurats de la tecnologia de la informació, que pretenen entendre les dades, seleccionar les rellevants, processar-les i crear un model vàlid amb resultats significatius.

3.2. Metodologies de mineria de dades

S'han proposat diverses metodologies pel desenvolupament de projectes de mineria de dades, tals com el SEMMA (*Sample, Explore, Modify, Model, Assess*), el DMAIC (*Define, Measure, Analyze, Improve, Control*) o el CRISP-DM (*Cross Industry Process for Data Mining*). Aquest darrer és el més utilitzat en els últims anys segons fonts informatives de *KD-Nuggets*, tal i com es pot observar a la *Figura 1*, i per aquest motiu s'han considerat les seves pautes de procediment per a la realització d'aquest treball.

3.2.1. CRISP-DM

El model CRISP-DM es basa en seguir de manera ordenada i rigorosa un total de sis fases diferents per dur a terme un anàlisi de dades. Aquestes es realitzen de manera cíclica per tal de millorar contínuament el model fins assolir un objectiu final, tal i com es pot veure representat al diagrama de flux a la *Figura 2*. A continuació es detallaran els passos a seguir per la metodologia.

Comprensió pel negoci

El primer pas a realitzar és l'avaluació de la situació inicial, on es defineixen els objectius i requisits del projecte des d'un punt de vista empresarial o institucional. Cal entendre bé quin és el problema a solucionar per no desviar-se de la finalitat de l'estudi. Això inclou saber de quines dades es disposa, què es vol predir i què es vol aconseguir amb la realització del projecte.

En aquesta fase es crea un full de ruta que determina el pla de desenvolupament del projecte. Tot i que el nom de la fase sembli destinat tan sols a negocis i projectes empresarials, també es pot aplicar a institucions sense ànim de lucre.

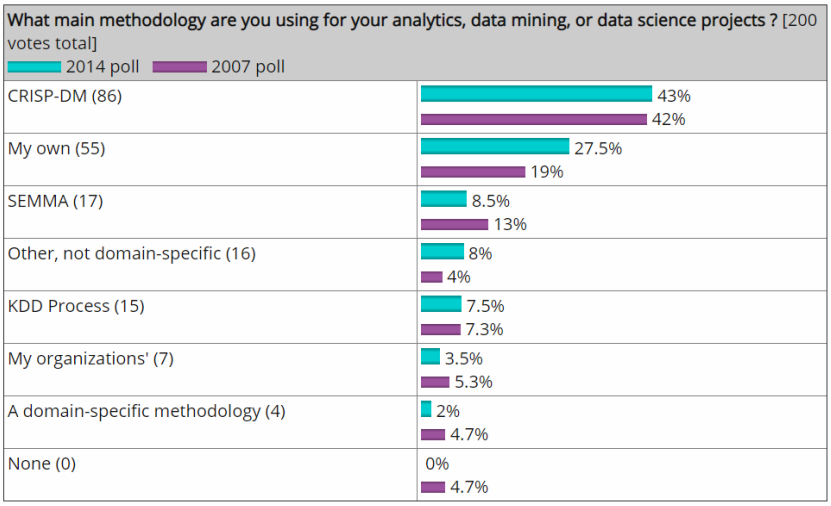


Figura 1: Taula comparativa dels models utilitzats en la mineria de dades, anys 2007-2014. Font d'informació: https://www.kdnuggets.com

Comprensió de dades

Un cop establerts els objectius empresarials i el pla de projecte, cal fer una recol·lecció inicial de dades i adaptar-les al seu imminent processament. Aquest pas pot incloure diferents taques, com ara descriure els camps de les dades, fer-ne una exploració mitjançant la visualització de gràfics estadístics o bé verificar la qualitat d'aquestes, com podria ser detectant valors nuls o fora de rang. En resum, s'ha d'avaluar la fiabilitat de les dades que es tractaran per tal de poder obtenir unes conclusions útils.

Preparació de dades

Una vegada seleccionades les dades a tractar durant el projecte, aquestes han de ser introduïdes al sistema operatiu en que seran tractades. En primer lloc s'explora quin és el programa informàtic més adequat pel projecte i s'importen les dades en el format necessari. Quan ja són introduïdes en el software, cal aplicar un *Data cleaning*, és a dir, realitzar les modificacions necessàries per poder tractar i processar adequadament les dades en la posterior creació del model. Finalment es procedeix a fer un *Data enrichment*, procés basat en enriquir el projecte amb noves dades que es creu que poden ser útils pel model, ja siguin de noves fonts o combinacions d'altres dades. **[3]**

Modelatge

El següent pas és escollir i construir un model adequat pels objectius que es volen assolir, tenint en compte la compatibilitat amb les dades que es disposa. De tots els models de predicció existents, la mineria de dades contempla bàsicament dues divisions, els models de regressió i els de classificació, dels que es parlarà amb més detall en els propers capítols. Aquesta fase inclou tenir en compte el mode en que serà avaluat el model per tal de poder validar els resultats obtinguts de manera fiable.

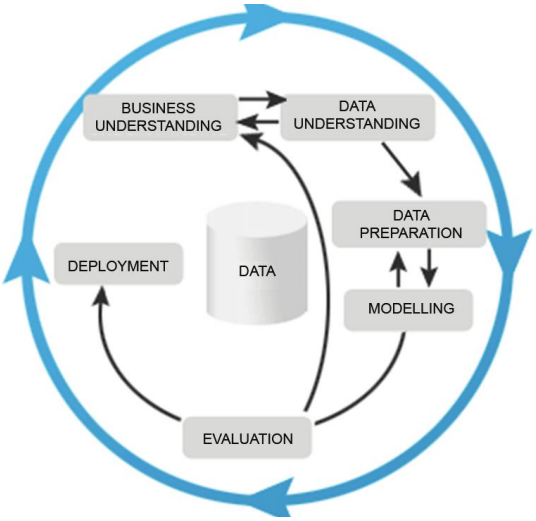


Figura 2: Diagrama de flux de la metodologia CRISP-DM

Avaluació

Els resultats obtinguts han de ser avaluats tenint en compte els objectius establerts a la primera fase. En aquesta fase entren en joc l'ús d'eines matemàtiques i estadístiques per poder fer una bona interpretació dels resultats obtinguts, parant especial atenció sobre quins paràmetres s'utilitzen per validar el model correctament. Aquest fet pot conduir a la identificació d'errors o modificacions que plantegin repetir fases anteriors del procediment o, en el pitjor dels casos, partir de zero com a un nou projecte. Sense una bona comprensió de dades no poden haver-hi resultats coherents amb els objectius.

Implementació

Després d'haver construït i validat el model, l'estudi realitzat proveeix un increment de coneixements que pot utilitzar-se per la finalitat que es desitgi, com seria el cas d'una presa de decisions per resoldre un problema. És aconsellable realitzar un informe on es documentin els resultats obtinguts de manera comprensible, ja que això permetrà enregistrar les evidències per a futurs estudis. Per altra banda, el model predictiu creat durant el projecte pot ser utilitzat com a eina de predicció per altres casos compatibles amb les seves restriccions.

Per a la realització d'aquest treball s'ha aplicat la metodologia CRISP-DM de manera parcial, adaptant les fases a la dimensió reduïda del projecte. S'ha procurat seguir la línia general de la metodologia per establir un ordre durant la realització del treball, així com poder tenir unes pautes de procediment pròpies d'un projecte de mineria de dades.

4. Eines utilitzades en el projecte

4.1. Llenguatge de programació

Una de les decisions més importants alhora d'iniciar un projecte de *Data Mining* és escollir el software informàtic en que es realitzarà el projecte sencer, incloent la construcció del model predicció i anàlisi de resultats. El programa escollit ha de complir un seguit d'especificacions, com poder modelar i manipular les dades, visualitzar-les, tractar-les o analitzar-les mitjançant eines matemàtiques i estadístiques de la complexitat desitjada.

Hi ha diversos llenguatges de programació que es poden utilitzar per a la ciència de dades, com ara el SQL, Java, Matlab, SAS, R i d'altres. Durant els darrers anys, el llenguatge R ha encapçalat la llista de programes més utilitzats a nivell mundial per la mineria de dades, gràcies a la gran capacitat de visualització de dades, el suport de la comunitat d'usuaris oferint paquets de documentació online i l'accessibilitat gratuïta per tot usuari.

No obstant això, la lentitud del programa davant de codis molts complexos i la difícil corba d'aprenentatge per l'usuari inexpert han donat pas a que *Python* ocupés el primer lloc a la llista de softwares més utilitzats per projectes *Data Mining*, tal i com podem veure a la *Figura 3* segons fonts de *KdNuggets*.

Per dur a terme aquest treball s'ha comptat amb el llenguatge *Python* com a eina de programació per realitzar tot el projecte. La decisió ha estat basada en els punts comentats anteriorment i impulsada pel fet de que és el programa amb més domini adquirit durant els anys de grau.

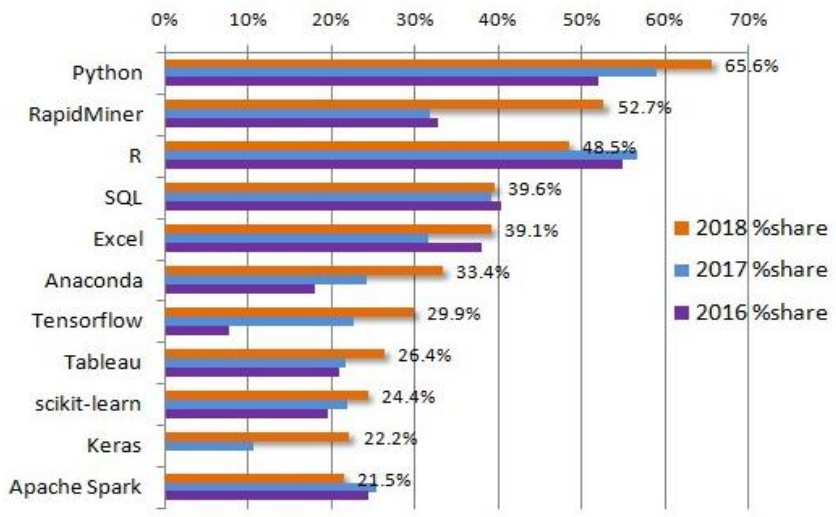


Figura 3: Gràfic comparatiu de softwares més utilitzats en la mineria de dades, anys 2016-2018. Font d'informació: https://www.kdnuggets.com

4.1.1. Llenguatge *Python* i plataforma *Anaconda*

Python és un popular llenguatge de programació orientat a objectes d'alt nivell i utilitzat àmpliament per un gran nombre de desenvolupadors de programari. Des del primer disseny creat per *Guido van Rossum* l'any 1991 la fundació ha anat desenvolupat aquest llenguatge de programació, focalitzant-se en la llegibilitat de codi i la computació científica i matemàtica. La sintaxi de *Python* és neta i curta, el seu llenguatge de codi obert es recolza sobre una biblioteca àmplia d'eines que enriqueixen el seu desenvolupament. A continuació es presentaran les seves característiques principals davant d'altres llenguatges de programació:

- Llenguatge robust i senzill per afavorir la corba d'aprenentatge de l'usuari.
- Compatibilitat amb una gran diversitat de plataformes, com ara Windows, Mac, Linux, etc.
- Llenguatge multi-paradigma que permet diversos estils com la programació orientada a objectes, la programació imperativa (mitjançant bucles) o la programació funcional (amb mòduls i funcions)

- Ús dinàmic de variables sense necessitat de definir el seu tipus prèviament (*int*, *float*, *string*, etc.)

Per la realització d'aquest treball s'ha comptat amb el suport de la plataforma *Anaconda*, una eina de distribució lliure i oberta per l'execució de llenguatge *Python*. Aquest programa disposa de diversos editors de codi, com ara el *Spyder*, proveïts amb una terminal que permet verificar el codi a mida que es va implementant, facilitant així la feina del programador. **[4]**

La plataforma permet el processament de grans volums d'informació, anàlisi predictiu i còmputs científics. Està orientada per simplificar el desplegament i administració dels paquets de software, fent-la una eina molt adient per la ciència de dades i aprenentatge automàtic (*Machine Learning*). Aquesta és utilitzada actualment per 6 milions d'usuaris i inclou més de 250 paquets de ciència de dades vàlides per a Windows, Linux i MacOS.

4.1.2. Llibreries del programa

La manipulació de dades s'utilitza per extreure, filtrar i transformar les dades de forma ràpida i senzilla amb un resultat eficient. Per aquesta finalitat, el programa *Python* disposa de diverses llibreries que ajuden a modelar les dades sense afegir complexitat al codi de programa. Les llibreries més utilitzades tant per l'usuari com per dur a terme aquest treball són les següents.

Pandas: És una potent llibreria d'anàlisi de dades que permet treballar amb taules de dades indexades mitjançant el mòdul *DataFrame*, permetent modificar o modelar qualsevol característica o valor d'aquestes. També proporciona eines d'estructures de dades com fusionar, modelar o tallar conjunts de dades. És molt eficaç, i permet carregar dades en format d'Excel, CSV, HDF5 i altres.

NumPy: Diminutiu de "*Python numèric"*, la llibreria *NumPy* s'utilitza en càlculs científics que proporcionen objectes de matriu, així com eines per integrar els llenguatges C i C++. *NumPy* proporciona una potent matriu *n* dimensional ordenada en files i columnes, les quals es poden exportar des d'una llista de *Python* o des d'una taula *DataFrame*.

Scikit-learn: Aquesta popular llibreria s'utilitza com a *Machine Learning* en ciències de dades amb diversos algorismes de classificació, regressió i agrupació. *Scikit* ofereix la possibilitat de crear diversos models de mineria de dades de manera senzilla i poder visualitzar la resposta, ja que està dissenyada per operar amb *SciPy, NumPy* i *Matplotlib*.

Matplotlib: Provinent del diminutiu de "*Mathematical Plotting Library*", s'utilitza principalment per a la visualització de dades, incloent gràfics 3D, histogrames, gràfics d'imatges, diagrames de dispersió i gràfics de barres. És funcional en gairebé totes les plataformes, com ara Windows, Mac i Linux. Aquesta llibreria també serveix com a extensió de *NumPy*, i disposa d'un mòdul *pyplot* que s'utilitza en les visualitzacions, sovint comparat amb els recursos del programa *Matlab*.

Aquest conjunt de llibreries han estat indispensables per la realització del treball, ja que han permès editar, modelar i afegir grans volums de dades de la manera desitjada en tot moment.

4.2. Models de predicció

De tots els models de predicció utilitzats en la mineria de dades se'n distingeixen bàsicament dos tipus, en funció de la resposta que es vol obtenir: el model de regressió i el de classificació. Per una banda, el model de regressió té com a objectiu predir valors continus, mentre que el de classificació s'utilitza per assignar una variable categòrica corresponent a una nombre limitat de classes.

Un bon exemple per entendre ambdós models seria la venda d'un pis mitjançant una empresa immobiliària. A partir de diferents variables del pis, com podrien ser els metres quadrats, la localització, el nombre d'habitacions, etc., es podrien crear dos models diferents. Per una banda, el model de classificació intentaria predir si l'empresa immobiliària estaria disposada a gestionar la venda del pis, és a dir, classificar la resposta en dos tipus: SÍ i NO. Per altre banda, el model de regressió treballaria per pronosticar el valor de venda en que es podria taxar el pis, en aquest cas una variable contínua expressada en milers d'euros.

Tornant a la realització d'aquest treball, cal recordar que l'objectiu és predir l'aprovat o suspès de les assignatures del tercer quadrimestre basant-se en les dades acadèmiques

prèvies de cada estudiant. Així doncs, es tracta clarament d'un problema de predicció per classificació, on la resposta adoptarà un nivell binari: Aprova o Suspèn. De tots els models disponibles s'ha optat per construir-ne dos molt utilitzats i suficientment diferents entre ells, la regressió logística i els arbres de decisió, comparant posteriorment les respostes obtingudes.

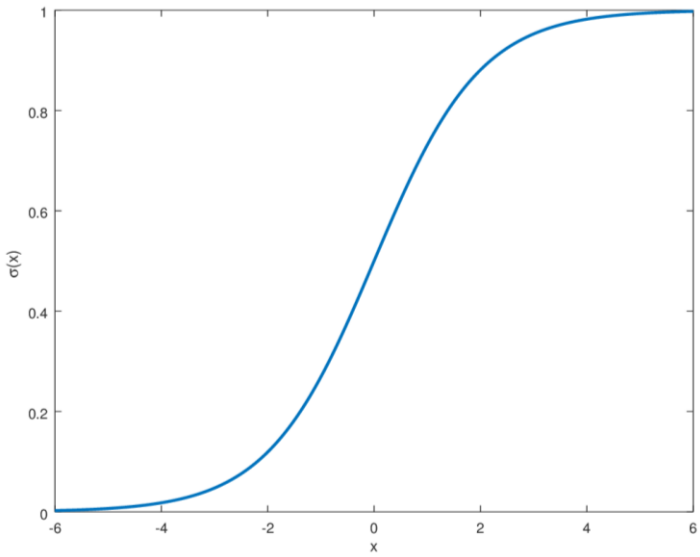
4.2.1. Regressió logística

Tot i que el seu nom pot induir a pensar el contrari, la regressió logística és una eina de classificació binària utilitzada per predir si es produirà o no un esdeveniment, senyalat amb el valor "1" si es produeix o "0" en cas contrari. El model treballa correlacionant diverses variables independents d'entrada X amb una única sortida binària Y corresponent a cada experiment. **[5]**

Per fer-ho, la regressió logística calcula la probabilitat de succés d'un o més experiments mitjançant la funció logística, també coneguda com a funció de *Sigmoid*. Aquesta funció pot representar la imatge de qualsevol nombre real a un valor comprès entre 0 i 1, ja que es tracta d'una probabilitat. Si el resultat es troba entre 0,5 i 1, la regressió logística categoritzarà la predicció com a esdeveniment produït "1", mentre que si el resultat es troba entre 0 i 0,5 es categoritzarà com a no produït "0". A la *Figura 4* es pot veure representada la funció logística, on la variable independent "X" és una combinació lineal de les variables d'entrada del model.

L'objectiu d'aquest model és utilitzar una font de dades d'entrenament per ajustar els valors dels coeficients β_n que minimitzin l'error entre el resultat previst i el resultat real. Un cop establerts els coeficients, el model ja estarà llest per poder predir nous casos.

Com es pot observar, la regressió logística és una eina adient pel cas que s'està estudiant, on es desitja predir un resultat binari (aprovat o suspès) per cada assignatura de segon curs. Així doncs, amb l'ajuda de la llibreria *scikit-learn*, s'ha utilitzat el mòdul *Logistic Regression* per construir un model de predicció basat en les variables independents de les que es disposa.



Funció logística:

$$\sigma(x) = \frac{1}{1+e^{-(x}}$$

Variables independents:

$$x = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Figura 4: Representacions gràfica i analítica de la funció logística

4.2.2. Arbre de decisions

El model de classificació d'arbre parteix d'una gran senzillesa i fàcil visualització de resultats. El seu funcionament es basa en la creació d'un arbre en el que un node arrel es va ramificant en funció de les decisions preses a cada nivell. Cada node intern de l'arbre correspon a un atribut, que mitjançat comparacions de valors acaba desembocant a un node final (o node de fulla), corresponent a una variable categòrica. A la *Figura 5* es pot veure un exemple d'arbre de decisions utilitzat en aquest treball. **[6]**

El principal repte de l'aplicació de l'arbre de decisions és identificar quins nodes s'han de considerar a cada nivell. Per resoldre aquest problema, s'aplica una funció matemàtica que atorga un valor a cada atribut segons el que ajuda a discriminar entre les classes a predir. Una de les funcions més utilitzades i amb la qual es realitzarà aquest treball és la impuresa de *Gini*, basada en mesurar la probabilitat en que un element escollit a l'atzar s'identifica com a incorrecte. Quan la impuresa de *Gini* és propera a 0, una gran quantitat de mostres del subconjunt de dades d'aquell node s'han categoritzat correctament, per tant serà un bon candidat a node final. Aquí entrarà en joc un nova variable, la quantitat mínima de mostres que es necessiten per poder dividir un node, ja que si les banques es separen massa ràpid es pot produir un error de model d'*Overfitting*, concepte que s'explicarà a continuació.

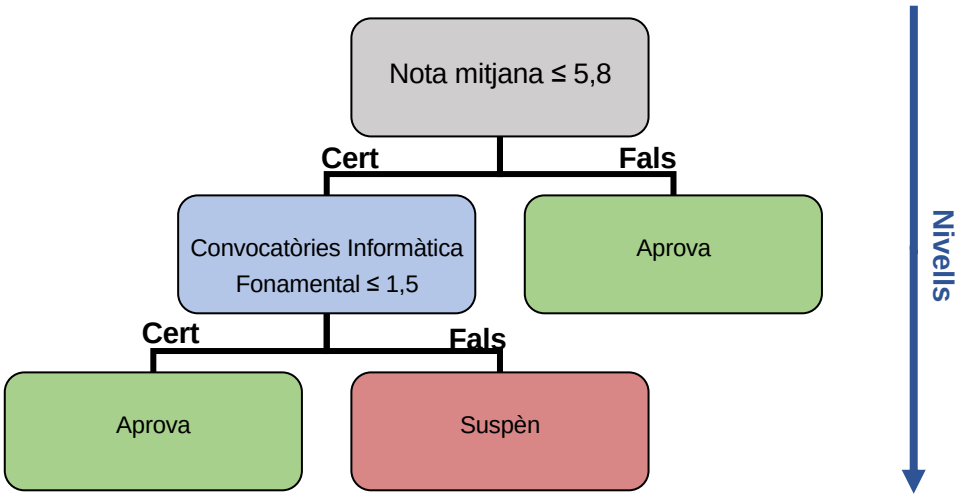


Figura 5: Representació d'un arbre de decisions emprat en el treball

4.2.2.1. *Overfitting*

Quan s'utilitzen eines de *Machine Learning* per construir models basats en valors reals sovint apareix un error de sobre-ajustament de resultats, conegut també com a *Overfitting*. Aquest fenomen és degut a causa d'intentar ajustar els resultats a un rang de valors massa ampli, on el model vol considerar inclosos aquells valors que es troben fora de la línia de tendència. Això condueix a una gran precisió pel model concret, però no representativa, doncs el model respondrà de manera imprecisa quan s'introdueixin unes dades diferents. A la *Figura 6* es pot veure com influeix el grau d'ajustament d'una corba a la predicció de la seva tendència.

L'error per sobre-ajustament és molt comú en els arbres de decisions, ja que si no s'estableix cap pauta durant la creació del l'arbre, aquest podria arribar a definir tants nodes finals com mostres analitzades, obtenint així una resposta correcta per cada mostra. Igual que les corbes d'exemples anteriors, l'arbre de decisions obtindria un rendiment perfecte però no representatiu, doncs seria pràcticament obsolet quan s'utilitzés amb unes altres dades. Per evitar aquest fet en la creació d'arbres de decisions es disposa de diferents recursos, generats per utilitzar menys decisions però amb més rellevància:

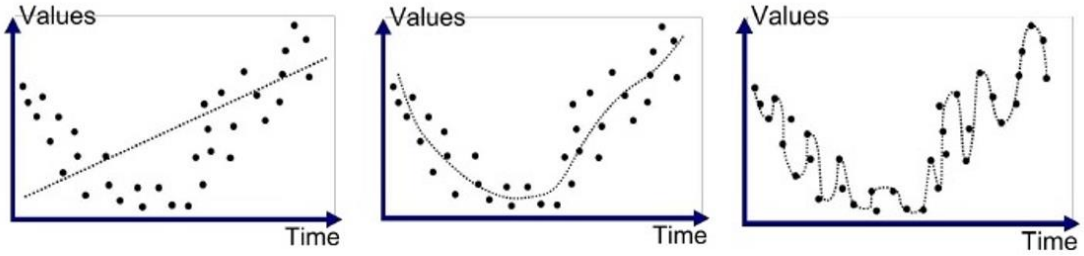


Figura 6: Exemples gràfics de corbes sub-ajustades (esquerra, Underfitting), ben ajustades (centre) i sobre-ajustades (dreta, Overfitting)

Max depth: Eina que permet establir un nombre límit de profunditat de nivells en un arbre.

Min samples leaf: Eina que permet establir el nombre mínim de mostres per definir un node final o node fulla.

Min samples split: Eina que permet establir el nombre mínim de mostres per poder dividir un node intern.

Max leaf nodes: Eina que permet establir el nombre màxim de nodes finals o nodes fulla.

Tots aquest recursos es troben disponibles a la classe *tree* de la llibreria *scikit-learn*, amb la que s'han configurat els arbres de decisions de cada assignatura a predir, construïts amb el mòdul *Decision Tree Classifier*. Jugant amb els valors dels diversos recursos esmentats s'ha pogut ajustar cada arbre per evitar l'*Overfitting*, obtenint així un millor rendiment del model.

4.3. Selecció de variables

El tret més important dels projectes de mineria de dades és identificar, categoritzar i atorgar valor a les variables que s'han considerat en el sistema d'estudi. Hom podria pensar que el millor procediment per avaluar dades es basa tan sols en introduir la quantitat més gran de dades disponibles i veure com respon el model. No obstant això no és sempre així, doncs hi ha moltes dades que poden generar factors de soroll que perjudiquin la resposta. Així doncs caldrà seleccionar les variables més rellevants i eliminar aquelles considerades

redundants, ja que quan menor sigui el nombre de variables més robust serà el model.

Per aquests motius esmentats, els analistes de dades han generat diversos mètodes i models de selecció de dades (en anglès *Feature selection*). D'aquests en destaquen els algoritmes meta-heurístics de selecció de variables, programats per localitzar un òptim global mitjançant processos combinatoris. En destaquen tres tipus principals:

Mètode de filtratge: Els mètodes de filtratge seleccionen variables basant-se només en característiques generals de correlació amb la variable a pronosticar. Funcionen suprimint les variables més redundants, utilitzant les demés pel posterior processat de modelatge. Aquests mètodes són particularment eficaços i resistents l'*overfitting*.

No obstant això, els mètodes de filtratge sovint poden seleccionar variables redundants, ja que no consideren les relacions entre variables independents, tan sols avaluen una per una la relació entre variable independent i variable objectiu.

Mètode "*wrapper*": Els mètodes *wrapper* avaluen subconjunts de variables utilitzant un mètode iteratiu que compara diferents mostres obtingudes per filtratge. D'aquesta manera s'intenta solucionar el problema de poder detectar les interaccions possibles entre variables. Tot i així, aquest mètode també disposa de certes desavantatges, com ara un risc creixent de l'*overfitting* quan el nombre d'observacions és insuficient, o bé un temps de computació significatiu quan s'eleva nombre de variables utilitzades.

Mètode encastat: Els mètodes encastats han estat recentment proposats per intentar combinar els avantatges dels dos mètodes anteriors. Un algoritme d'aprenentatge realitza la selecció i la classificació simultàniament, iterant repetides vegades el procediment fins obtenir el millor resultat. Com és d'imaginar, aquest mètode és d'elevada complexitat i necessita ajustar-se adequadament segons l'objecte d'estudi.

Per la realització d'aquest treball, els models de selecció de variables han anat lligats de la mà dels models de predicció, ja que s'ha anat veient com responien els resultats de predicció segons les variables seleccionades i el model de predicció utilitzat. Així doncs, el mètode que millor s'adaptava al procediment iteratiu esmentat era un model de selecció de variables per filtratge.

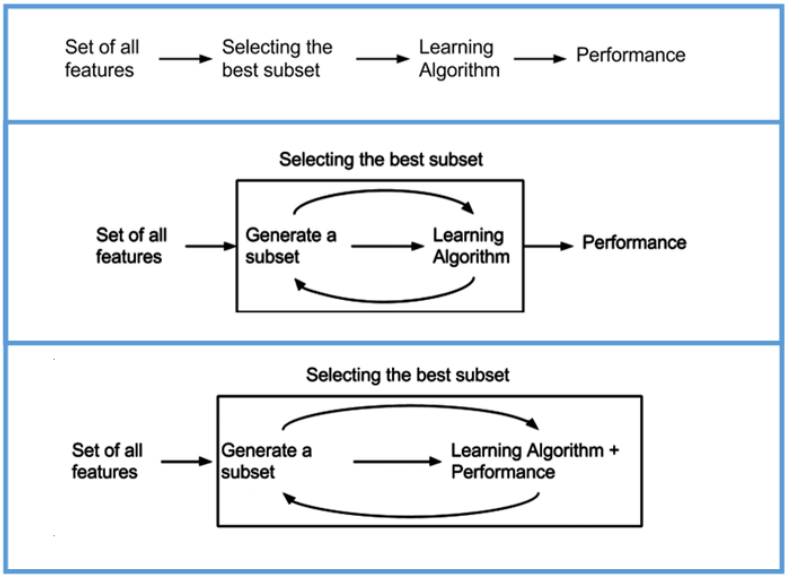


Figura 7: Funcionament gràfic dels tres tipus de selecció de variables: Per filtratge (amunt), per wrapper (centre) i per encastament (a sota)

Tanmateix el llenguatge *Python* ofereix una classe anomenada *SelectKBest*, altre vegada provinent de la llibreria *scikit-learn*, que selecciona per filtratge les variables que tenen una relació més directa amb el resultat de sortida. La classe funciona d'una manera senzilla, analitza per separat cada variable independent atorgant-li un valor K en funció de la relació amb el valor de sortida, calculat mitjançant una funció estadística de prova escollida. Posteriorment selecciona el nombre de variables desitjat segons ordre decreixent del valor K. La funció de relació escollida ha estat *xi quadrat*, ja que és una de les més utilitzades per models de classificació de variables amb valors no negatius. **[7]**

4.4. Validació de resultats i indicadors de rendiment

Una part fonamental del treball és l'anàlisi i validació de resultats obtinguts a partir de tots els passos comentats anteriorment. Per fer-ho, és important fer una bona comprensió tant dels paràmetres a avaluar com del rang de valors que han de prendre per ser considerats com a bons. En el món de l'anàlisi estadístic sovint es cometen errors greus de confusió alhora d'interpretar resultats, que poden conduir a falses conclusions contraproduents per l'objectiu d'estudi. Per tal d'evitar aquest fet s'intentarà explicar de manera clara els indicadors utilitzats per l'avaluació de resultats.

En primer lloc cal entendre com es realitza un estudi de predicció de resultats en mineria de dades. Un cop es disposa de les dades ben preparades per a l'execució dels models de predicció, aquestes es separen en dos subconjunts: dades d'entrenament "*Train*" i de prova "*Test*". De cada subconjunt es separen les variables independents d'entrada "X" de les respectives respostes que s'han obtingut "Y" per cada experiment. Així doncs es disposa de quatre subconjunts de variables: "X_train", "Y_train", "X_test", "Y_test".

Tant els models de predicció com els de selecció de variables es configuren a partir de les dades "Train" per tal de no rebre influències dels valors utilitzats per l'avaluació, doncs això afavoriria falsament als posteriors resultats. Els models es posen a prova amb les dades "Test", on es crea una nova variable corresponent als resultats predits "Y_pred" segons les entrades a predir "X_test". Aquesta nova variable s'utilitza per comparar el valor predit "Y_pred" envers el valor real "Y_test" que hauria de prendre cada experiment. D'aquí s'extreuen indicadors com la matriu de confusions o la corba ROC.

4.4.1. Matriu de confusions

La matriu de confusions és una eina d'avaluació de resultats per models de predicció classificatoris, on la resposta tan sols pot prendre un valor categòric corresponent a un nombre limitat de classes. La seva funcionalitat es basa en exposar les coincidències trobades entre els valors predits i els valors reals, creant així una matriu amb les diferents possibilitats. A continuació s'explicaran els elements que la composen, tenint en compte que en aquest cas tan sols hi ha una resposta binària. **[8]**

True Positive (TP): Nombre de dades que s'han predit com a valor "1" i han coincidit amb el valor real.

True Negative (TN): N Nombre de dades que s'han predit com a valor "0" i han coincidit amb el valor real.

False Positive (FP): Nombre de dades que s'han predit com a valor "1" i no han coincidit amb el valor real, en aquest cas corresponent a un valor "0".

False Negative (FN): Nombre de dades que s'han predit com a valor "0" i no han coincidit amb el valor real, en aquest cas corresponent a un valor "1".

		Valor predit	
		0	1
Valor real	0	TN	FP
	1	FN	TP

Figura 8: Elements de la matriu de confusions

L'objectiu del treball és predir el aprovat o suspès de cada assignatura de segon. Com que hi ha més aprovat que suspesos, l'estudi s'ha centrat més en predir els suspesos, ja que és el valor més difícil de pronosticar degut a que és el menys freqüent. Així doncs, s'ha establert el valor "1" com a "suspès" i "0" com a "aprovat".

A partir dels elements de la matriu de confusions se'n poden extreure diferents indicadors per poder avaluar les relacions de coincidència entre valors predits i valors encertats. A continuació es detallaran els més utilitzats per projectes de mineria de dades, i els que han servit com a base del projecte realitzat.

Precisió: Percentatge d'encerts totals del model de predicció. De tots els valors predits, quants d'aquests s'han encertat.

$$Precisió = \frac{TP + TN}{TP + TN + FP + FN}$$

Exactitud: Percentatge d'encerts de la predicció de cada classe. De tots els valors predits per cada classe "0" i "1", quants d'aquests s'han encertat.

$$Exactitud\,(P) = \frac{TP}{TP + FP} \qquad Exactitud\,(N) = \frac{TN}{TN + FN}$$

Sensibilitat: També conegut com a *Recall*, és la proporció de casos positius "1" que han estat predits pel model de predicció. De tots els valors reals "1" que hi havia, quants d'aquests s'han pogut predir.

$$Sensibilitat = \frac{TP}{TP + FN}$$

Especificitat: Proporció de casos negatius “0” que han estat predits pel model de predicció. De tots els valors reals “0” que hi havia, quants d’aquests s’han pogut predir.

$$Especificitat = \frac{TN}{TN + FP}$$

Com és d’imaginar, els indicadors explicats són relativament proporcionals al rendiment del model de precisió, és a dir, quan més elevats siguin aquests índexs més robust serà el model. No obstant això, s’ha de ser conscient de les diferències entre els indicadors, doncs una bona precisió tampoc assegura que el model sigui fiable.

Per exemple, imaginem una suposada font de dades amb un 80% de valors positius “1” i un 20% restant de valors negatius “0”. Imaginem que el model predicció tan sols retorna valors positius, sense pronosticar mai cap valor negatiu. Aquest model estaria obtenint un 80% de precisió, un 80% d’exactitud de positius i un 100% de sensibilitat. Tot i semblar uns resultats excel·lents, el model tan sols es limita a retornar el mateix valor independentment de les variables, i per tant tindria un 0% tant d’especificitat com d’exactitud de negatius.

Per tal d’agrupar els indicadors esmentats, els analistes de dades es recolzen sobre una eina gràfica coneguda com a corba ROC, que té en compte la relació entre els paràmetres comentats.

4.4.2. Corba ROC

La corba *Reciever Operating Characteristic* és una representació gràfica de la sensibilitat enfront l'especificitat per a un sistema classificador binari, traçada segons avança el llindar de discriminació. Els eixos que componen la corba són l’índex de TP (sensibilitat) a l’eix de les ordenades, i l’índex de FP (1 - especificitat) a l’eix de les abscisses. Aquesta corba proporciona eines tant visuals com analítiques per avaluar un model de predicció binari.

La corba es traça mitjançant la unió de diferents punts d’estudi obtinguts a partir d’una mostra acumulativa que té en compte cada cop més valors, fins arribar a la totalitat de les dades. De cada mostra es representen els dos índexs respecte el conjunt total com a punt a l’espai ROC, de tal manera que sempre s’inicia el traçat a l’origen de coordenades i es finalitza al punt (1,1). En altres paraules, el traçat comença a l’origen (0,0) i avança en direcció vertical si l’experiment ha estat encertat o en direcció horitzontal en cas contrari.

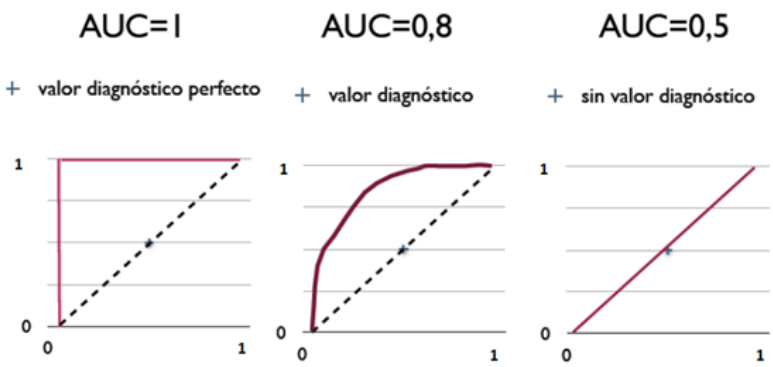


Figura 9: Gràfics de corbes ROC: Model ideal (esquerra), model bo (centre), model aleatori (dreta)

Aquesta corba és una bona eina visual ja que ràpidament es pot veure si el traçat es mou per l'hemisferi de bona predicció o bé es manté en la diagonal de predicció aleatòria, com podria ser els resultats del llançament d'una moneda. També es pot obtenir un indicador molt usat a nivell d'anàlisi, l'àrea compresa sota la corba o paràmetre AUC (*Area Under Curve*). Aquest darrer indicador mostra una relació percentual que engloba els encerts de classe positiva i els de negativa, on l'àrea sempre serà compresa entre 0 i 1.

4.4.3. Correlació entre variables independents

Tenint en compte els objectius principals d'aquest treball, un exercici molt important alhora d'avaluar resultats serà identificar la rellevància de les variables seleccionades en els models de predicció. Més enllà de la influència individual de cada variable sobre la resposta, també s'haurà de tenir en compte la relació entre les variables independents utilitzades en el model. No seran d'interès aquelles que mantinguin alta proporcionalitat entre elles, ja que no aporten un augment d'informació i poden generar soroll en la resposta, perjudicant el rendiment del model. Així doncs, interessarà obtenir variables el més independents possible per poder abordar el problema des de diferents perspectives.

Un bon indicador per percebre la relació entre variables és el coeficient de correlació. La correlació lineal indica la relació de proporcionalitat entre dues variables estadístiques quantitatives. Diem que hi ha correlació entre variables quan els valors d'una d'elles varia sistemàticament respecte els valors homònims de l'altre, és a dir, quan els valors

d'ambdues variables augmenten o disminueixen de manera proporcional.

Existeixen diferents coeficients de correlació en funció de la naturalesa de les dades a tractar. En aquest treball s'utilitzarà el coeficient de correlació de *Pearson*, ja que és el més utilitzat i representatiu del càlcul estadístic. Aquest estableix una relació entre la covariància i la desviació estàndard de cada variable, obtenint un valor comprès entre -1 i 1. Els valors propers a 1 indicaran una correlació positiva en que la dependència de les variables actua de manera proporcional, mentre que els propers a -1 seran indicadors d'una correlació negativa de proporcional inversa entre variables. Així doncs, l'objectiu serà obtenir un coeficient de correlació proper a 0, que suposarà una bona independència entre variables.

Aquest indicador serà emprat durant l'etapa de validació de resultats per saber si les variables seleccionades tenen relació de dependència entre elles. En cas que afirmatiu, s'avaluarà la opció de repetir l'experiment prescindint d'una de les variables dependents per veure si aquesta influïa negativament en la resposta. D'aquesta manera es pretén polir el model de predicció perquè estigui constituït tan sols de variables significatives.

5. Implementació del projecte

Un cop definides les diferents eines utilitzades en el projecte es procedeix a explicar el transcurs experimental, els passos que s'han seguit per dur a terme la seva realització d'inici a fi. Tractant-se d'un projecte de mineria de dades, s'ha intentat seguir la línia de la metodologia CRISP-DM de forma adaptada a les dimensions i objectius del projecte. Aquest fet comporta aplicar un ordre durant la realització de les accions pràctiques del projecte, començant per una fase de comprensió i preparació de dades, una posterior creació del model i una etapa de validació de resultats, que pot incloure repetir els passos previs cíclicament. Després d'avaluar el model final es procedirà a debatre sobre la fase d'implementació del projecte.

5.1. Comprensió i preparació de dades

El projecte parteix d'una sèrie de dades proveïdes per l'escola, les quals engloben informació dels estudiants i estudiantes matriculats a l'ETSEIB durant el període de temps comprès entre els anys 2010 i 2017. Les dades es troben desglossades en tres fulls de càlcul d'*Excel*, corresponents a les dades personals, dades acadèmiques de la fase inicial i dades acadèmiques de la fase no inicial de cada alumne enregistrat. Cal esmentar que el treball realitzat s'ha basat tan sols en aquesta font d'informació, en cap moment s'han considerat dades que no fossin aquestes ni s'ha procedit a fer una nova recol·lecció.

Tan bon punt s'ha disposat de les dades dels estudiants, els tres fulls de càlcul d'*Excel* s'han importat com a tres respectius objectes *DataFrame* de la llibreria *Pandas*, amb els quals s'ha realitzat la resta de treball mitjançant el llenguatge *Python*. Per fer més amena la feina de programació s'ha comptat amb la ajuda de l'editor *Spyder*, el qual ha facilitat aquesta labor permetent comprovar les accions realitzades a mida que es va escribint el llenguatge de codi.

Treballant ara amb el llenguatge de programació, s'ha procedit a fer una primera visualització de dades per comprendre cada aspecte de la informació disposada. En primer lloc, s'ha pogut veure com els tres objectes *DataFrame* s'estructuren d'una manera molt similar. Cada fila conté informació d'un estudiant concret i cada columna el valor d'una variable d'estudi. En particular es distingeixen dues estructures diferents; les dades perso-

Variable	Definició	Tipus	Exemple
CODI_EXPEDIENT	Número d'expedient de l'estudiant	*int64*	226431
SEXE	H gènere masculí, D gènere femení	*object*	D
CP_FAMILIAR	Codi postal del domicili de l'estudiant	*object*	*08029
ANY_ACCES	Any de la primera matrícula	*int64*	2012
TIPUS_ACCES	1 si s'accedeix per les PAU	*int64*	1
NOTA_ACCES	Nota obtinguda a les PAU	*float64*	10,346
CP_CENTRE_SEC	Codi postal del centre del centre d'estudis previs	*float64*	*08021

Taula 1: Descripció de variables de les dades personals

Variable	Definició	Tipus	Exemple
CODI_PROGRAMA	Codi del programa d'estudis	*int64*	752
CODI_EXPEDIENT	Número d'expedient de l'estudiant	*int64*	226431
CODI_UPC_UD	Codi de l'assignatura	*int64*	240011
CREDITS	Crèdits de l'assignatura	*float64*	7,5
CURS	Any en que s'ha cursat l'assignatura	*int64*	2013
QUAD	Quadrimestre de tardor (1) o primavera (2)	*int64*	2
SUPERA	S si aprova l'assignatura, N en cas contrari	*object*	S
NOTA_PROF	Nota final de l'assignatura	*float64*	4,8
NOTA_NUM_AVAL	Nota més alta obtinguda a l'assignatura	*float64*	4,8
NOTA_NUM_DEF	Nota final representada a l'expedient	*float64*	5,0
GRUP_CLASSE	Grup matriculat d'aquella assignatura	*object*	41

Taula 2: Descripció de variables de les dades acadèmiques

nals, formades per una única fila d'informació per estudiant, i les dades acadèmiques, on cada fila informa sobre una convocatòria per estudiant i assignatura. Per exemple, si un estudiant ha repetit una assignatura, la informació de les dues convocatòries es representa en dues files independents, tot i que facin referència al mateix estudiant.

El significat de cadascuna de les variables s'ha representat en les *Taules 1* i *2*, corresponents a les dades personals i les dades acadèmiques, doncs l'estructura de les fases inicial i no inicial és exactament la mateixa.

En una primera instància s'ha decidit començar a treballar tan sols amb les dades acadèmiques, contemplant la possibilitat d'ampliar posteriorment la informació amb les dades personals dels estudiants. Aquesta decisió s'ha pres per comprovar el grau d'influència d'unes notes sobre unes altres.

Tractant ara només amb les dades acadèmiques, el següent pas a realitzar ha estat transformar-les per tal de poder treballar posteriorment amb elles. Aquesta fase s'ha focalitzat en crear una nova estructura que englobi la informació resumida de cada estudiant procedent d'ambdues dades acadèmiques, les de la fase inicial i les de la no inicial. Per fer-ho possible s'han realitzat diverses accions essencials en el tractament de dades.

5.1.1. Filtratge

Donat que el treball realitzat tan sols contempla els estudiants del Grau en enginyeria en tecnologies industrials de l'ETSEIB, la primera tasca ha estat identificar aquest tipus d'individus i separar-los dels demés. Aquesta tasca és de rellevant importància, ja que per una banda s'està definint la població de l'estudi i per l'altra s'estan evitant posteriors problemes. Els estudiants d'altres graus de l'ETSEIB cursen assignatures diferents, així que incloure'ls a l'estudi augmentaria la variabilitat de resultats de rendiment acadèmic.

La tasca de filtratge ha estat relativament senzilla gràcies a la variable CODI_PROGRAMA que informa del grau que està cursant l'estudiant. Així doncs tan sols ha calgut seleccionar els estudiants amb el codi 752 pertanyent al grau d'enginyeria industrial. Aquesta fase també s'ha aprofitat per eliminar els possibles valors repetits erròniament.

5.1.2. Pivotatge de files i columnes

El pivotatge de dades consisteix en endreçar files i columnes d'un objecte de dades (en aquest cas un objecte *DataFrame*) de la manera desitjada. Una decisió important a prendre durant aquesta fase ha estat estructurar les dades per poder resumir la informació de cada estudiant. Donat que es desitja partir del model més bàsic i enriquir-lo posteriorment, s'ha decidit treballar amb el valor de la variable NOTA_NUM_DEF de cada assignatura. Aquesta variable es representa amb una columna on cada element correspon a la nota obtinguda per un estudiant en una convocatòria concreta d'una assignatura.

L'objectiu plantejat es basa en establir un nou ordre de files i columnes. Es planteja resumir tota la informació d'un estudiant en una sola fila, i representar la informació de cada assignatura cursada en una columna independent, tal i com es pot veure a la *Taula 3*. Així doncs, de cada fila actual interessa el valor de NOTA_NUM_DEF obtingut per un estudiant CODI_EXPEDIENT corresponent a una convocatòria de l'assignatura CODI_UPC_UD.

Aquí sorgeix el primer problema a solucionar, donat que existeix un gran variabilitat de possibilitats diferents segons el nombre d'assignatures cursades per estudiant. És a dir, hi ha estudiants que hauran cursat tan sols un cop cada assignatura i n'hi ha que n'hauran cursat més degut a les assignatures que s'han repetit. Aquest fet significa prendre una decisió sobre com estructurar les dades, ja que es desitja obtenir una estructura uniforme on cada estudiant disposi del mateix nombre de variables.

El problema exposat s'ha solucionat de manera diferent en les dades de la fase inicial respecte les de la fase no inicial. En primer lloc, cada assignatura de la fase inicial s'ha desglossat en dues variables; la nota obtinguda a la última convocatòria i el nombre de convocatòries realitzades en aquella assignatura. Per les dades de la fase no inicial, s'ha aplicat una funció que retorna un 0 si l'alumne ha aprovat l'assignatura en la primera convocatòria o un 1 en cas contrari.

D'aquesta manera tots els estudiants disposen d'exactament dues variables per cada assignatura de la fase inicial i una variable per cada assignatura de la fase inicial. Mitjançant el pla d'estudis de la pròpia escola s'ha pogut identificar cada assignatura a partir de la variable CODI_UPC_UD, modificant-la com a un *string* amb el nom de l'assignatura corresponent.

CODI_EXPEDIENT	CODI_UPC_UD	NOTA_NUM_DEF
226410	240012	5,6
226410	240013	8,3
226431	240012	5,2
226431	240013	6,1

CODI_EXPEDIENT	240012	240013
226410	5,6	8,3
226431	5,2	6,1

Taula 3: Primera transformació de les dades acadèmiques

5.1.3. Concatenació i eliminació de valors buits

El següent pas a realitzar és identificar i unir les files d'estudiants de les dades de la fase inicial amb les de la no inicial. En aquest cas s'ha comptat amb la ajuda de la funció de concatenació *join* de *Python*, que permet fusionar les dues fonts d'informació en un únic objecte *DataFrame* utilitzant com a clau d'unió el CODI_EXPEDIENT de cada estudiant. La realització d'aquest pas s'aprofita per eliminar dos casos possibles: el cas de que un estudiant aparegui a una base de dades i no en l'altre degut a algun possible error d'enregistrament de dades, i el cas de que apareguin valors buits degut a que un estudiant hagi abandonat els estudis o provingui d'un altre grau convalidant algunes assignatures.

Així doncs s'ha obtingut el resultat final de les dades a tractar per la creació del model, on cada estudiant definit pel CODI_EXPEDIENT disposa de de 26 variables: 10 variables de la última nota obtinguda a cada assignatura del primer curs, 10 variables corresponents al nombre de convocatòries de cadascuna d'aquestes assignatures i 6 variables binàries que indiquen si l'estudiant ha aprovat (0) o suspès (1) la primera convocatòria de cadascuna de les assignatures del tercer quadrimestre. La *Taula 4* mostra una part visual de l'estructura final descrita, composta per un total de 2.282 estudiants

5.2. Selecció de variables i creació de models

Amb les dades filtrades, netejades i transformades s'ha pogut iniciar el procés de selecció automàtica de variables i construcció de models de predicció (regressió logística i arbre de

CODI_EXPEDIENT	Nota Àlgebra	Convocatòria Àlgebra	Nota Càlcul 1	…	Mecànica	Informàtica
226410	6,6	1	7,2	…	0	0
226431	6,2	2	5,8	…	1	0

Taula 4: Exemple de l'estructura final de les dades transformades

decisions). Aquests procediments han estat realitzats de forma independent per cada variable a predir, per tant, s'han creat sis parells de models de predicció amb sis respectius conjunts de dades, un per cada assignatura del tercer quadrimestre.

En primer lloc s'han definit els models de predicció mitjançant les classes *Logistic Regression* pel model de regressió logística i *Decision Tree Classifier* per l'arbre de decisions, ambdues classes provinents de la llibreria *scikit-learn*. Cadascun dels sis parells de models necessita les seves dades corresponents; les variables independents X i la variable dependent Y corresponent a l'assignatura a predir, tal i com es pot veure a la *Taula 5*. Els models es creen en base a un subconjunt de mostres d'entrenament *Train* i es validen amb les mostres restants *Test*. En aquest cas s'ha pres com a mostres *Train* un 75% del subconjunt de dades de cada model. Les mostres han estat seleccionades de manera aleatòria per tal d'evitar possibles coincidències entre dades.

Acte seguit s'ha aplicat el mètode de selecció de variables per filtratge fent ús la classe *SelectKBest* de la llibreria *scikit-learn*. Fixant explícitament un nombre n, aquesta classe retorna el subconjunt de n variables independents X que afecten més a la resposta Y segons un test estadístic sobre valors categòrics. El nombre n s'ha augmentat progressivament des d'1 fins a 20 per observar l'ordre de rellevància de variables que atorga aquesta classe, tal i com es pot veure a la *Taula 6*.

D'aquesta manera s'ha pogut provar la resposta dels sis parells de models segons el nombre n de variables independents considerades. És a dir, cada model creat s'ha posat a prova primer amb la variable independent X més significativa, després amb les dues més significatives, després amb tres, i així successivament fins a utilitzar totes les variables.

CODI_EXPEDIENT	Nota Àlgebra	Convocatòria Àlgebra	Nota Càlcul 1	…	Materials
226410	6,6	1	7,2	…	0
226431	6,2	2	5,8	…	1
…	…	…	…	…	…
226889	7,5	1	8,2	…	0
227006	5,0	3	5,6	…	1
	20 variables independents X				1 variable dependent Y

Taula 5: Exemple del conjunt de les dades utilitzades pels models de predicció de l'assignatura de Materials, indicant les dades Train (taronja) i les Test (verd)

n	1	2	3	4	5	…
Variable que s'afegeix	Nota Informàtica Fon	Convocatòria Informàtica Fon	Convocatòria Termodinàmica	Convocatòria Mecànica Fon	Convocatòria Geometria	…

Taula 6: Exemple d'ordre de rellevància de les variables que afecten a la predicció de l'aprovat o suspès de l'assignatura d'Informàtica

5.3. Selecció de variables i creació de models

Com bé s'ha comentat, de cada assignatura s'ha hagut de modificar el model de predicció fins a vint vegades, una per cada nova variable significativa afegida a les dades X. Per obtenir uns resultats coherents i comparables, la aleatorietat de les dades *Train* i *Test* ha estat la mateixa pels vint experiments de cada model, és a dir, s'ha definit arbitràriament una llista d'estudiants *Train* i *Test* i s'ha mantingut durant tots els experiments. Això ha estat possible fixant el valor del generador d'aleatorietat *random state*.

De cada experiment s'han recollit les dades dels diferents indicadors de resultats comentats en el capítol anterior. Cal recordar que ens referirem als valors positius 1 com a Suspès i als valors negatius 0 com a Aprovat, ja que l'objectiu d'estudi és predir els suspesos, donat

que aquests són la minoria.

- **TN:** Resultat predit com a Aprovat (0) i encertat.
- **FP:** Resultat predit com a Suspès (1) i fallat.
- **FN:** Resultat predit com a Aprovat (0) i fallat.
- **TP:** Resultat predit com a Suspès (1) i encertat.
- **Precisió:** Percentatge de resultats predits correctament, tan aprovats com suspesos.
- **Exactitud (0):** Percentatge de resultats predits com a aprovats (0) i encertats.
- **Exactitud (1):** Percentatge de resultats predits com a suspesos (1) i encertats.
- **Especificitat (0):** Percentatge d'aprovats reals trobats pel model de predicció.
- **Sensibilitat (1):** Percentatge de suspesos reals trobats pel model de predicció.
- **AUC:** Àrea compresa sota la corba ROC, relacionant especificitat i sensibilitat.
- **Variable entrant:** Nova variable afegida al model de predicció.

També s'han pogut utilitzar eines gràfiques proveïdes pel propi llenguatge de programació *Python*, com corbes ROC, representacions d'arbres de decisió o matrius de confusió de cada experiment realitzat. Aquestes eines han permès percebre errors i valors a anòmals a simple vista, així com avaluar el rendiment dels models creats de manera ràpida i efectiva. Es pot veure un exemple a les *figures 10* i *11*.

No obstant, cal remarcar que aquestes eines gràfiques han estat usades com a informació complementària d'ajuda mentre es procedia a la creació dels models. Això és degut a que proveeixen informació d'un experiment concret, cosa que faria molt laboriosa la feina d'anàlisi de resultats si es té en compte l'elevat nombre d'experiments a realitzar de cada model. Per aquest fet, s'ha decidit realitzar l'anàlisi i validació dels models en base a taules de resultats numèrics recollits, com podrien ser els indicadors percentuals descrits anteriorment.

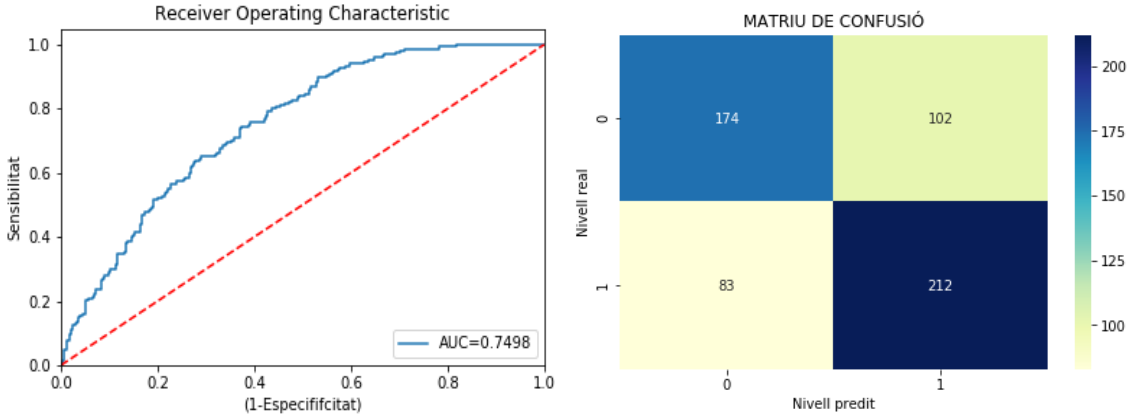


Figura 10: Corba ROC (esquerra) i matriu de confusió (dreta) del model de predicció de regressió logística per l'assignatura de Mecànica per amb n=5

A diferència de la regressió logística, el model d'arbre de decisió necessita un ajust previ per tal d'evitar l'*Overfitting* comentat en capítols anteriors. Mitjançant les variables *max depth*, *min samples leaf* i *max leaf nodes* s'ha pogut ajustar cada arbre de decisió per obtenir uns resultats fiables. El valor d'aquestes variables d'ajustament ha vingut condicionat pel nombre n de variables utilitzades a cada experiment. A major nombre de variables independents, més nodes s'han hagut d'incloure per poder veure reflectit aquest nou increment d'informació.

Finalment, per tal de garantir uns resultats més fiables, cada model (amb els seus vint experiments) s'ha repetit fins a trenta vegades i s'ha treballat amb la mitjana aritmètica dels resultats obtinguts. Aquest valor s'ha decidit en funció del temps d'execució del programa, doncs a partir d'aquest nombre de repeticions el programa generava un retard molt significatiu. S'ha fixat una generador d'aleatorietat fix durant cada repetició, de manera que cada repetició disposa de les mateixes dades *Train* i les mateixes dades *Test* com bé s'ha comentat anteriorment. Cada repetició disposa del seu generador d'aleatorietat particular definit amb l'eina *random state*.

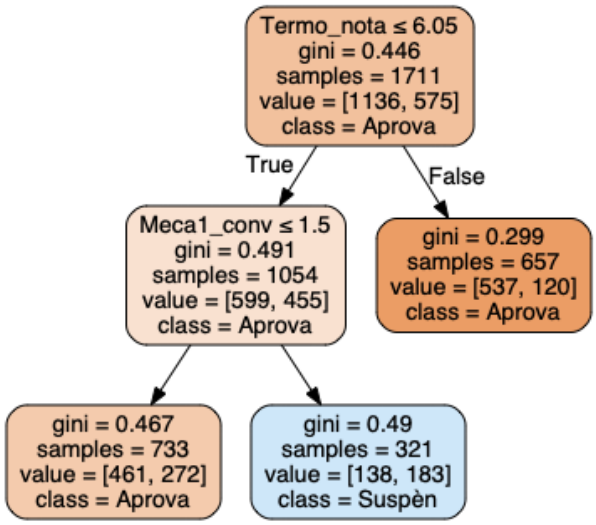


Figura 11: Arbre de decisions de l'assignatura d'Electromagnetisme amb n=2

5.3.1. Anàlisi de resultats de l'assignatura de Mètodes Numèrics

La primera assignatura a avaluar ha estat la de Mètodes Numèrics, així que s'ha procedit a recopilar els resultats obtinguts dels models de predicció en dues taules, una pel model de regressió logística (*Taula 7*) i l'altra pel d'arbre de decisions (*Taula 8*).

A primera vista ambdós models semblen molt fiables ja que presenten uns resultats de precisió entorn al 87%, un valor significatiu molt alt. Aquest fet significa que si s'apliqués qualsevol dels dos models de predicció amb individus aleatoris que just han acabat la fase inicial, s'encertarien entre 8 i 9 de cada 10 resultats pronosticats. No obstant, no tan sols s'ha d'observar aquesta dada, hi ha altres factors determinants pel rendiment dels models.

Seguint amb els indicadors percentuals, un tret molt característic i sospitós dels resultats obtinguts són els valors d'especificitat i sensibilitat, on ambdós models han respòs amb un 99% i un 0% respectivament. Aquest fet indica que els models han pronosticat l'aprovat de l'assignatura sigui quin sigui el valor de la variable independent, tal i com indiquen els baixos valors de FP i TP dels models per aquest cas.

Regressió logística de l'assignatura de Mètodes Numèrics												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**…**	**20**
TN	497,50	495,70	494,47	493,70	494,00	494,33	493,73	493,73	493,70	493,67	…	492,87
FP	0,27	2,07	3,30	4,07	3,77	3,43	4,03	4,03	4,07	4,10	…	4,90
FN	72,03	70,30	69,63	69,17	69,17	69,23	68,00	67,90	67,50	67,73	…	68,13
TP	1,20	2,93	3,60	4,07	4,07	4,00	5,23	5,33	5,73	5,50	…	5,10
Precisió	0,8734	0,8733	0,8723	0,8717	0,8723	0,8727	0,8738	0,874	0,8747	0,8742	…	0,8721
Exactitud (0)	0,8735	0,8758	0,8766	0,8771	0,8772	0,8772	0,8789	0,8791	0,8797	0,8794	…	0,8785
Exactitud (1)	0,6778	0,6213	0,5445	0,5045	0,5274	0,5597	0,5783	0,5834	0,5981	0,5866	…	0,5204
Especificitat (0)	0,9995	0,9959	0,9934	0,9918	0,9924	0,9931	0,9919	0,9919	0,9918	0,9918	…	0,9901
Sensibilitat (1)	0,0163	0,0408	0,0494	0,0557	0,0558	0,0548	0,0712	0,0726	0,0787	0,0757	…	0,0696
AUC	0,6252	0,6657	0,6882	0,6927	0,7285	0,7267	0,7324	0,7326	0,7377	0,7369	…	0,7431
Variable entra	Meca1 conv	Info1 conv	Termo conv	Geom conv	Quim2 nota	Calc2 conv	Quim2 conv	Algeb conv	Calc1 nota	Calc1 conv	…	Termo nota

Taula 7: Resultats de la regressió logística de l'assignatura de Mètodes Numèrics

Arbre de decisions de l'assignatura de Mètodes Numèrics												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**…**	**20**
TN	497,77	497,20	497,10	497,10	494,87	494,63	492,20	492,10	492,27	491,40	…	489,37
FP	0,00	0,57	0,67	0,67	2,90	3,13	5,57	5,67	5,50	6,37	…	8,40
FN	73,23	72,97	73,00	73,00	71,90	71,83	70,17	70,20	70,20	69,93	…	69,63
TP	0,00	0,27	0,23	0,23	1,33	1,40	3,07	3,03	3,03	3,30	…	3,60
Precisió	0,8717	0,8712	0,871	0,871	0,869	0,8687	0,8674	0,8671	0,8674	0,8664	…	0,8633
Exactitud (0)	0,8717	0,8721	0,872	0,872	0,8732	0,8732	0,8753	0,8752	0,8752	0,8755	…	0,8755
Exactitud (1)	0,0000	0,0107	0,0172	0,0172	0,1041	0,1012	0,1867	0,1832	0,1765	0,1984	…	0,2269
Especificitat (0)	1,0000	0,9989	0,9987	0,9987	0,9942	0,9937	0,9888	0,9886	0,9889	0,9872	…	0,9831
Sensibilitat (1)	0,0000	0,0043	0,0035	0,0035	0,0191	0,0197	0,0415	0,0410	0,0405	0,0446	…	0,0498
AUC	0,625	0,6574	0,684	0,6797	0,6918	0,6885	0,6932	0,6915	0,681	0,6783	…	0,6766
Variable entra	Meca1 conv	Info1 conv	Termo conv	Geom conv	Quim2 nota	Calc2 conv	Quim2 conv	Algeb conv	Calc1 nota	Calc1 conv	…	Termo nota

Taula 8: Resultats de l'arbre de decisions de l'assignatura de Mètodes Numèrics

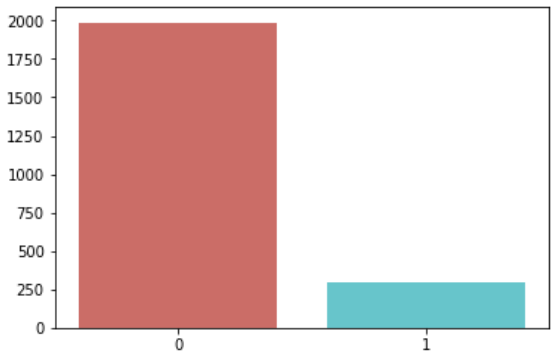


Figura 12: Diagrama de barres dels aprovats (0) i suspesos (1) de la variable Mètodes Numèrics

Aquesta observació condueix a fer un estudi de la variable a tractar, en aquest cas, l'aprovat o suspès de l'assignatura d'Informàtica. Fent un simple diagrama de barres dels valors de la variable dependent (*Figura 12*), s'ha pogut veure que els aprovats d'aquesta assignatura superen als suspesos sumant un 87,1% dels alumnes totals matriculats en la primera convocatòria. De 2.282 alumnes matriculats, tan sols 295 han suspès l'assignatura el primer cop que la cursaven.

D'aquesta manera s'explica la gran diferència percentual d'especificitat envers sensibilitat dels models, ja que el suspès real serà molt difícil de trobar degut a la seva poca densitat de resultats. Com a conseqüència d'aquest fet es pot observar que la relació d'àrea AUC no és gaire alta fins que no s'augmenta considerablement el nombre de variables n.

Un altre fet que pot haver influït en la resposta el model és la informació adquirida de les variables independents. Es pot observar que el model es basa tan sols en el nombre de convocatòries en set de les primeres vuit variables considerades. Aquest fet pot haver fet més difícil quantificar la relació amb la variable dependent, ja que no es té en compte cap variable referent a les notes.

Pel que fa als models de predicció, es pot veure una tendència molt similar en ambdós casos, tot i que la regressió logística supera considerablement a l'arbre de decisions en els valors en l'exactitud d'aprovats. En aquest cas, es mostra sorprenentment un pic en d'exactitud d'aprovats amb només una variable independent, la convocatòria de Mecànica Fonamental. Per aquest motiu, la solució mes factible per la predicció de l'assignatura de Mètodes Numèrics seria utilitzant aquesta variable en el model de regressió logística.

D'aquesta manera la predicció d'aprovats i suspesos seria encertada entorn al 87% i el 68% dels casos respectivament, uns valors significativament alts. No obstant, els suspesos reals tan sols s'haurien pronosticat com a tals en el 4% dels casos. Aquest fet indica que el model de predicció tan sols atorga la classificació de suspès quan hi ha un alt índex de probabilitat de que succeeixi.

5.3.2. Anàlisi de resultats de l'assignatura d'Informàtica

Pel que fa a l'assignatura d'Informàtica, s'han resumit els resultats obtinguts dels models de predicció en les *Taules* 9 i *10*. El tret més rellevant que es pot observar a simple vista és la selecció de variables dels models, on s'ha considerat com a variables més significatives la nota i la convocatòria de l'assignatura precedent: Informàtica Fonamental. Aquest fet confirma intuïcions mitjançant el mètode de selecció de variables.

Els indicadors percentuals també són força alts en aquest cas, començant per uns valors de precisió entorn al 80% dels resultats pronosticats, un valor significativament elevat. No obstant, altre vegada podem observar uns resultats sospitosos d'especificitat i sensibilitat, on ambdós models han respòs amb un 100% i un 0% respectivament per l'experiment d'una única variable (n=1).

De nou es torna a fer un anàlisi exploratori de dades de la variable dependent a predir, ja que es sospita un cas similar a l'assignatura de Mètodes Numèrics. En aquest cas són 470 alumnes de 2.282 els que han suspès la primera convocatòria de l'assignatura d'informàtica, representant així un 20,6% del total. Així doncs, es tracta d'un cas semblant d'excés d'aprovats envers els suspesos.

De totes maneres, en comparació a l'assignatura comentada anteriorment, els models de l'assignatura d'Informàtica han obtingut uns millors resultats en la predicció suspesos. Aquesta afirmació és deguda a que s'ha vist un increment de sensibilitat entorn al 20%, així com uns valors més estables d'exactitud de suspesos.

Pel que fa als models de predicció, ambdós presenten resultats molt similars en tots els indicadors. El model de regressió logística obté uns valors més alts d'exactitud de suspesos i del paràmetre AUC, mentre que l'arbre de decisions ho fa en la sensibilitat, el qual té molta importància en aquest cas. Així doncs es buscarà un punt d'equilibri entre els dos models.

Regressió logística de l'assignatura d'Informàtica												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	453,73	446,97	440,30	440,47	438,33	437,80	437,23	436,63	436,53	436,10	...	433,60
FP	0,00	6,77	13,43	13,27	15,40	15,93	16,50	17,10	17,20	17,63	...	20,13
FN	117,27	108,10	97,67	96,53	95,30	95,23	93,33	93,33	91,57	92,77	...	93,23
TP	0,00	9,17	19,60	20,73	21,97	22,03	23,93	23,93	25,70	24,50	...	24,03
Precisió	0,7946	0,7988	0,8054	0,8077	0,8061	0,8053	0,8076	0,8066	0,8095	0,8067	...	0,8015
Exactitud (0)	0,7946	0,8053	0,8185	0,8203	0,8215	0,8214	0,8241	0,8239	0,8266	0,8247	...	0,8231
Exactitud (1)	0,0000	0,5970	0,5996	0,6127	0,5916	0,5808	0,5932	0,5851	0,5987	0,5826	...	0,5458
Especificitat (0)	1,0000	0,9851	0,9704	0,9708	0,9661	0,9649	0,9637	0,9623	0,9621	0,9612	...	0,9556
Sensibilitat (1)	0,0000	0,0786	0,1672	0,1771	0,1877	0,1881	0,2044	0,2046	0,2193	0,2094	...	0,2057
AUC	0,6859	0,7201	0,7351	0,7415	0,7464	0,756	0,7575	0,7581	0,758	0,7616	...	0,7632
Variable que entra	Info1 nota	Info1 conv	Termo conv	Meca1 conv	Geom conv	Expre nota	Expre conv	Quim1 nota	Calc1 conv	Quim2 nota	...	Termo nota

Taula 9: Resultats de la regressió logística de l'assignatura d'Informàtica

Arbre de decisions de l'assignatura d'Informàtica												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	453,73	441,30	442,13	436,07	435,83	433,73	431,70	430,33	431,13	429,60	...	422,50
FP	0,00	12,43	11,60	17,67	17,90	20,00	22,03	23,40	22,60	24,13	...	31,23
FN	117,27	107,03	100,87	93,27	93,47	92,77	93,57	93,27	92,13	91,43	...	88,47
TP	0,00	10,23	16,40	24,00	23,80	24,50	23,70	24,00	25,13	25,83	...	28,80
Precisió	0,7946	0,7908	0,803	0,8057	0,805	0,8025	0,7975	0,7957	0,7991	0,7976	...	0,7904
Exactitud (0)	0,7946	0,8049	0,8143	0,8239	0,8235	0,824	0,8221	0,8221	0,8242	0,8247	...	0,8269
Exactitud (1)	0,0000	0,3910	0,6101	0,5785	0,5724	0,5535	0,5162	0,5043	0,5232	0,5192	...	0,4836
Especificitat (0)	1,0000	0,9726	0,9745	0,9610	0,9605	0,9560	0,9515	0,9485	0,9502	0,9468	...	0,9311
Sensibilitat (1)	0,0000	0,0874	0,1399	0,2042	0,2023	0,2090	0,2027	0,2051	0,2148	0,2206	...	0,2448
AUC	0,6743	0,7054	0,7182	0,7308	0,7259	0,7377	0,7279	0,7192	0,725	0,7181	...	0,7094
Variable que entra	Info1 Nota	Info1 conv	Termo conv	Meca1 conv	Geom conv	Expre nota	Expre conv	Quim1 nota	Calc1 conv	Quim2 nota	...	Termo nota

Taula 10: Resultats de l'arbre de decisions de l'assignatura d'Informàtica

Posant la mirada a la selecció automàtica de variables, en el model de classificació d'arbre de decisions es pot veure com apareix un pic relatiu en n=4 variables, ja que amb el següent valor de n disminueixen tots els indicadors percentuals. Aquest fet pot indicar que les següents variables tan sols aporten soroll a la resposta, així que es podria fixar un model de predicció basant-se tan sols en quatre variables: convocatòries d'Informàtica Fonamental, Mecànica Fonamenta i Termodinàmica Fonamental i última nota d'Informàtica Fonamental.

D'aquesta manera, el model de selecció d'arbre podria predir amb èxit el 82% dels aprovats i entorn al 58% dels suspesos, havent trobat el 96% dels aprovats reals i al voltant del 20% dels suspesos reals. Es pot veure que en aquest cas es redueix l'exactitud de la predicció de suspesos per augmentar la sensibilitat, en altres paraules, el model arrisca més alhora de pronosticar un valor positiu (suspès).

5.3.3. Anàlisi de resultats de l'assignatura d'Electromagnetisme

Per l'assignatura d'Electromagnetisme s'han resumit els resultats obtinguts dels models de predicció en les *Taules 11* i *12*, referents a la regressió logística i l'arbre de decisió respectivament. De nou, no és d'estranyar que les dues primeres variables seleccionades pel model facin referència a les assignatures de Termodinàmica i Mecànica Fonamental, doncs ambdues pertanyen al Departament de Física de l'ETSEIB, al igual que l'assignatura d'Electromagnetisme.

En relació als indicadors percentuals, es pot veure un augment significatiu de la sensibilitat en comparació a les assignatures estudiades recentment, doncs en alguns casos s'ha arribat a valors propers al 45%. També es pot veure un anivellament de l'exactitud dels resultats, obtenint uns valors més baixos pel cas d'aprovats (a prop d'un 74%) i uns més alts pels suspesos (entorn al 60%) respecte les assignatures anteriors.

Aquesta observació condueix a pensar que l'assignatura es suspèn més que les anteriors comentades, ja que està costant més predir l'aprovat però s'està incrementant la predicció de suspesos reals. Fent de nou un estudi dels valors obtinguts per la variable dependent s'ha pogut veure com el nombre de suspesos en la primera convocatòria d'Electromagnetisme augmentava fins al 33,6% sumant un total de 767 estudiants dels 2.282 enregistrats.

Regressió logística de l'assignatura d'Electromagnetisme												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	373,63	340,33	330,00	331,37	329,07	324,43	323,97	322,50	322,37	322,07	...	321,20
FP	3,53	36,83	47,17	45,80	48,10	52,73	53,20	54,67	54,80	55,10	...	55,97
FN	191,60	141,43	128,43	123,07	118,63	113,07	112,53	113,07	111,93	108,67	...	106,20
TP	2,23	52,40	65,40	70,77	75,20	80,77	81,30	80,77	81,90	85,17	...	87,63
Precisió	0,6583	0,6878	0,6925	0,7043	0,708	0,7096	0,7097	0,7062	0,708	0,7132	...	0,716
Exactitud (0)	0,6617	0,7064	0,7199	0,7291	0,7351	0,7417	0,7423	0,7406	0,7424	0,7479	...	0,7518
Exactitud (1)	0,0129	0,5869	0,5826	0,6079	0,6113	0,6058	0,6056	0,5976	0,6008	0,6088	...	0,611
Especificitat (0)	0,9911	0,9023	0,8750	0,8786	0,8725	0,8603	0,8591	0,8552	0,8549	0,8541	...	0,8518
Sensibilitat (1)	0,0127	0,2701	0,3374	0,3651	0,3884	0,4171	0,4199	0,4171	0,4230	0,4398	...	0,4525
AUC	0,6565	0,6992	0,7259	0,7382	0,7411	0,7552	0,7552	0,756	0,7619	0,7631	...	0,7713
Variable entra	Termo nota	Meca1 conv	Quim2 nota	Calc2 conv	Termo conv	Meca1 nota	Geom conv	Calc1 nota	Geom nota	Calc1 conv	...	Expre conv

Taula 11: Resultats de la regressió logística de l'assignatura d'Electromagnetisme

Arbre de decisions de l'assignatura d'Electromagnetisme												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	377,17	333,77	313,83	326,80	322,93	305,53	306,43	307,23	306,93	306,80	...	307,77
FP	0,00	43,40	63,33	50,37	54,23	71,63	70,73	69,93	70,23	70,37	...	69,40
FN	193,83	136,47	123,47	130,20	128,47	117,97	117,53	119,07	116,80	117,00	...	113,40
TP	0,00	57,37	70,37	63,63	65,37	75,87	76,30	74,77	77,03	76,83	...	80,43
Precisió	0,6605	0,685	0,6729	0,6838	0,68	0,668	0,6703	0,669	0,6724	0,6719	...	0,6799
Exactitud (0)	0,6605	0,7099	0,7192	0,7156	0,7156	0,7221	0,7234	0,7214	0,725	0,7245	...	0,7311
Exactitud (1)	0,0000	0,5685	0,5346	0,5647	0,5491	0,5174	0,5216	0,5181	0,5240	0,5228	...	0,5386
Especificitat (0)	1,0000	0,8847	0,8322	0,8665	0,8561	0,8102	0,8126	0,8147	0,8140	0,8136	...	0,8161
Sensibilitat (1)	0,0000	0,2952	0,3636	0,3288	0,3371	0,3920	0,3943	0,3862	0,3982	0,3970	...	0,4155
AUC	0,6506	0,6971	0,705	0,7075	0,7064	0,7114	0,7134	0,7138	0,7128	0,7111	...	0,7125
Variable entra	Termo nota	Meca1 conv	Quim2 nota	Calc2 conv	Termo conv	Meca1 nota	Geom conv	Calc1 nota	Geom nota	Calc1 conv	...	Expre conv

Taula 12: Resultats de l'arbre de decisions de l'assignatura d'Electromagnetisme

En referència als models de predicció, ambdós models no discrepen de més d'un 8% en cap dels seus indicadors per cada experiment. No obstant, la regressió logística obté uns resultats lleugerament superiors a la majoria de camps d'estudi, de manera que no hi ha lloc a dubte en l'elecció del model de predicció.

Pel que fa a la selecció de variables, si bé és cert que l'ordre de prioritats segueix la línia de la intuïció, sembla no haver-hi suficient relació per definir el model amb poques variables. El model es comença a estabilitzar passades les 10 primeres variables significatives, cosa que indica no haver-ne trobat una especialment rellevant envers a les demés.

Així doncs, el model final de predicció vindrà definit per una regressió logística composta d'un mínim de 10 variables. Es podrà predir correctament a prop d'un 75% dels aprovats i un 61% dels suspesos, trobant el 85% dels aprovats reals i el 44% dels suspesos reals. Tot i no ser els millors resultats de precisió trobats fins el moment, aquest model presenta una gran relació de predicció d'aprovats i suspesos, que és l'objectiu del projecte. Com a prova d'aquest fet s'obté un alt paràmetre AUC (76%), que estableix una bona relació entre sensibilitat i especificitat.

5.3.4. Anàlisi de resultats de l'assignatura de Mecànica

Pel que fa a l'assignatura de Mecànica, s'han resumit els resultats obtinguts dels models de regressió logística i d'arbre de decisió en les *Taules 13* i *14* respectivament.

El tret més característic observat a simple vista és que per primera vegada s'ha obtingut un valor més gran del 50% a tots els indicadors percentuals de la taula, superant el el 60% en la majoria de casos. Per aquest motiu s'ha decidit explorar i visualitzar les dades corresponents a la variable dependent a predir. A la *Figura 13* es pot veure clarament que més de la meitat d'estudiants no superen l'assignatura el primer cop que la cursen, concretament un 50,8% del total.

Aquest fet explica el comportament dels resultats obtinguts pels models de predicció. Degut a que ens trobem davant d'un cas de resultat binari equitatiu, els indicadors percentuals s'han anivellat significativament entre ells, obtenint uns valors pròxims al 65% com a tendència general. Tot i no ser uns resultats brillants, aquest fet garanteix que els models pronostiquin correctament un mínim de dues terceres parts en qualsevol tipus de predicció de resultats.

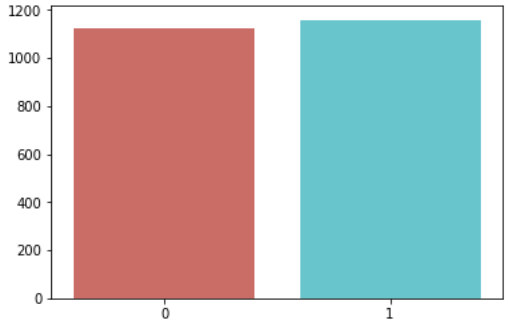


Figura 13: Diagrama de barres dels aprovats (0) i suspesos (1) de la variable Mecànica

Aquest fet explica el comportament dels resultats obtinguts pels models de predicció. Degut a que ens trobem davant d'un resultat binari equitatiu, els indicadors percentuals s'han anivellat significativament entre ells, obtenint uns valors pròxims al 65% com a tendència general. Tot i no ser uns resultats brillants, aquest fet garanteix que els models pronostiquin correctament un mínim de dues terceres parts en qualsevol tipus de predicció de resultats.

En relació als dos tipus de models de predicció, la regressió logística obté uns millors resultats de precisió d'aprovats i suspesos. Tot i que l'arbre de decisions obté una especificitat lleugerament superior al model de regressió logística, aquesta obté una sensibilitat considerablement més alta. Per aquest fet, també s'obtenen uns valors de AUC més alts per aquest model, així que la balança es decantarà per utilitzar de regressió logística com a model de predicció.

Pel que fa a la selecció automàtica de variables, s'ha pogut localitzar un parell de pics relatius en les dades procedents d'ambdós models quan s'ha utilitzat n=4 i n=7 variables independents. A partir d'aquest nombre de variables el sistema s'estabilitza o inclòs comença a decréixer de valor dels indicadors, tant a la variable següent n+1 com en el cas d'utilitzar les vint variables. Donat que els resultats són pràcticament iguals pels dos valors de n, es seleccionarà el nombre més baix de variables per tal de fer el model més eficient.

Així doncs, utilitzant quatre variables significatives, de tots els casos predits com a aprovats i suspesos s'hauran encertat un 70% i un 67% respectivament. Un 60% dels aprovats reals de l'assignatura hauran estat ben localitzats pel model, així com un 76% dels suspesos. Aquestes dades confirmen que el model pot ser utilitzat amb un cert índex de fiabilitat, doncs obté un rendiment superior a la predicció aleatòria (50%) en tots els casos possibles.

Regressió logística de l'assignatura de Mecànica												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	160,53	162,40	165,70	166,47	166,97	166,77	170,47	170,60	170,87	170,47	...	176,53
FP	116,77	114,90	111,60	110,83	110,33	110,53	106,83	106,70	106,43	106,83	...	100,77
FN	81,23	76,97	75,07	69,87	70,77	70,20	71,13	72,47	72,07	72,63	...	75,90
TP	212,47	216,73	218,63	223,83	222,93	223,50	222,57	221,23	221,63	221,07	...	217,80
Precisió	0,6532	0,664	0,6731	0,6835	0,6828	0,6835	0,6883	0,6862	0,6874	0,6857	...	0,6906
Exactitud (0)	0,6649	0,6789	0,6889	0,7046	0,7027	0,7042	0,7057	0,7021	0,7035	0,7014	...	0,6995
Exactitud (1)	0,6457	0,6537	0,6623	0,6692	0,6691	0,6694	0,6759	0,6749	0,6758	0,6744	...	0,6839
Especificitat (0)	0,5793	0,5858	0,5978	0,6008	0,6024	0,6018	0,6151	0,6156	0,6165	0,6151	...	0,6369
Sensibilitat (1)	0,7239	0,7382	0,7448	0,7623	0,7592	0,7610	0,7578	0,7533	0,7547	0,7528	...	0,7417
AUC	0,7002	0,7268	0,7319	0,7429	0,7461	0,7525	0,7553	0,7583	0,7577	0,7571	...	0,7623
Variable entra	Termo nota	Algeb nota	Calc1 nota	Meca1 nota	Quim2 nota	Calc2 nota	Info1 conv	Termo conv	Geom nota	Info1 nota	...	Quim1 conv

Taula 13: Resultats de la regressió logística de l'assignatura de Mecànica

Arbre de decisions de l'assignatura de Mecànica												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	146,67	167,37	163,87	179,90	176,10	169,97	169,60	172,63	173,27	176,27	...	176,50
FP	130,63	109,93	113,43	97,40	101,20	107,33	107,70	104,67	104,03	101,03	...	100,80
FN	68,30	90,80	81,53	96,70	93,43	90,03	91,07	94,00	98,10	99,93	...	100,77
TP	225,40	202,90	212,17	197,00	200,27	203,67	202,63	199,70	195,60	193,77	...	192,93
Precisió	0,6516	0,6485	0,6586	0,6601	0,6591	0,6543	0,6519	0,6521	0,646	0,648	...	0,647
Exactitud (0)	0,6844	0,6539	0,6712	0,6514	0,6547	0,6545	0,6512	0,6482	0,6402	0,6402	...	0,6375
Exactitud (1)	0,6337	0,6509	0,6527	0,6702	0,6654	0,6558	0,6540	0,6572	0,6539	0,6585	...	0,6571
Especificitat (0)	0,5286	0,6040	0,5914	0,6488	0,6354	0,6132	0,6120	0,6229	0,6254	0,6363	...	0,6368
Sensibilitat (1)	0,7672	0,6913	0,7228	0,6708	0,6821	0,6938	0,6904	0,6805	0,6668	0,6606	...	0,6569
AUC	0,6883	0,7074	0,7117	0,7185	0,7207	0,7166	0,7156	0,7132	0,7101	0,7091	...	0,7086
Variable entra	Termo nota	Algeb nota	Calc1 nota	Meca1 nota	Quim2 nota	Calc2 nota	Info1 conv	Termo conv	Geom nota	Info1 nota	...	Quim1 conv

Taula 14: Resultats de l'arbre de decisions de l'assignatura de Mecànica

5.3.5. Anàlisi de resultats de l'assignatura de Materials

Per l'assignatura de Materials s'han executat ambdós models de predicció, recollint els resultats obtinguts en dues noves taules: *Taula 15* i *Taula 16*. A primera vista s'observa que les dues primeres assignatures determinades com a significatives són les notes de Química 1 i Química 2. Aquest fet no és d'estranyar, doncs les bases de la ciència de materials recauen en el món de la química.

Parant especial atenció als indicadors percentuals d'aquest cas, es pot veure un màxim relatiu en n=5 pel cas de l'arbre de decisions, ja que absolutament tots els seus valors decreixen en prendre una nova variable d'estudi (n=6). Aquest fet indica que la següent variable entrant (nota d'Expressió Gràfica) podria estar perjudicant la resposta del model. Per altra banda confirma la rellevància de les assignatures de química i termodinàmica seleccionades fins el moment, incloent també la nota de Mecànica Fonamental.

Posant en comparació els dos models de classificació es pot veure com altra vegada els valors obtinguts per la regressió logística són lleugerament superiors als de l'arbre de decisions. Per aquest motiu s'ha decidit seleccionar aquest model per la predicció de l'aprovat i suspès de l'assignatura de Materials. En aquest cas s'ha pogut observar que amb n=7 variables s'obtenen uns valors d'indicadors percentuals superiors a qualsevol altre experiment realitzat amb un nombre de variables superior.

D'aquesta manera es podria obtenir un model de predicció per regressió logística amb un 74% d'encerts en els pronòstics d'aprovat de Materials i fins un 56% en el cas dels suspesos. Pel que fa a la localització de valors reals, s'obtindria una especificitat del 89% i una sensibilitat del 31%. Aquests resultats no són especialment dolents, tot i que el model queda descentrat en la localització d'aprovats i suspesos de l'assignatura.

Per comprovar aquesta última afirmació s'ha realitzat de nou un estudi de la variable dependent, on s'ha pogut veure que el 68,8% dels estudiants aprova l'assignatura el primer cop que la cursa. Aquest fet condueix a rebaixar el rendiment de la predicció d'aprovats per augmentar la de suspesos, tot i que en aquest cas no queda ben compensat, fent que el model no sigui fiable ni per una cosa ni per l'altre. Prova d'això és l'escàs valor del paràmetre AUC en ambdós models, ja que aquest és l'indicador que relaciona especificitat i sensibilitat.

Regressió logística de l'assignatura de Materials												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	367,00	355,30	354,37	352,27	348,93	348,30	347,77	347,60	344,53	344,20	...	341,87
FP	24,40	36,10	37,03	39,13	42,47	43,10	43,63	43,80	46,87	47,20	...	49,53
FN	157,73	146,40	136,40	132,60	128,70	126,80	123,67	123,60	125,60	125,37	...	122,63
TP	21,87	33,20	43,20	47,00	50,90	52,80	55,93	56,00	54,00	54,23	...	56,97
Precisió	0,681	0,6804	0,6963	0,6992	0,7002	0,7025	0,7070	0,7068	0,698	0,6978	...	0,6985
Exactitud (0)	0,701	0,7085	0,7223	0,7267	0,7307	0,7333	0,7378	0,7379	0,733	0,7332	...	0,7362
Exactitud (1)	0,2833	0,4882	0,5423	0,5472	0,5477	0,5518	0,5635	0,5632	0,5367	0,5357	...	0,5364
Especificitat (0)	0,9383	0,9081	0,9057	0,9002	0,8916	0,8900	0,8886	0,8883	0,8804	0,8795	...	0,8736
Sensibilitat (1)	0,1249	0,1857	0,2415	0,2625	0,2841	0,2945	0,3120	0,3125	0,3013	0,3024	...	0,3176
AUC	0,6714	0,6894	0,7034	0,7129	0,7187	0,7249	0,7303	0,7294	0,7298	0,7294	...	0,7305
Variable entra	Quim2 nota	Quim1 nota	Termo conv	Meca1 conv	Termo nota	Expre nota	Meca1 nota	Info1 nota	Geom conv	Info1 conv	...	Expre conv

Taula 15: Resultats de la regressió logística de l'assignatura de Materials

Arbre de decisions de l'assignatura de Materials												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	387,23	360,77	362,57	347,67	338,73	334,93	333,60	334,07	332,20	332,87	...	326,43
FP	4,17	30,63	28,83	43,73	52,67	56,47	57,80	57,33	59,20	58,53	...	64,97
FN	176,73	154,63	146,77	135,60	129,33	129,47	126,57	126,93	125,53	125,50	...	122,57
TP	2,87	24,97	32,83	44,00	50,27	50,13	53,03	52,67	54,07	54,10	...	57,03
Precisió	0,6832	0,6755	0,6925	0,6859	0,6813	0,6744	0,6771	0,6773	0,6765	0,6777	...	0,6716
Exactitud (0)	0,6871	0,702	0,7127	0,7202	0,7243	0,7218	0,7252	0,7251	0,726	0,7266	...	0,7274
Exactitud (1)	0,0272	0,3294	0,5534	0,5164	0,4910	0,4754	0,4833	0,4813	0,4790	0,4824	...	0,4703
Especificitat (0)	0,9896	0,9224	0,9268	0,8886	0,8657	0,8560	0,8523	0,8537	0,8489	0,8506	...	0,8343
Sensibilitat (1)	0,0167	0,1413	0,1848	0,2465	0,2811	0,2803	0,2954	0,2937	0,3009	0,3012	...	0,3186
AUC	0,6550	0,6733	0,6865	0,6881	0,6831	0,6723	0,6819	0,6797	0,6779	0,6767	...	0,681
Variable entra	Quim2 nota	Quim1 nota	Termo conv	Meca1 conv	Termo nota	Expre nota	Meca1 nota	Info1 nota	Geom conv	Info1 conv	...	Expre conv

Taula 16: Resultats de l'arbre de decisions de l'assignatura de Materials

5.3.6. Anàlisi de resultats de l'assignatura d'Equacions Diferencials

Finalment, l'última assignatura a analitzar és la d'Equacions diferencials, d'on s'ha plasmat la resposta dels models de predicció creats en les *Taules 17* i *18*. Pel que fa a la selecció de variables, es pot veure que la variable independent més significativa és la nota de Càlcul 1, una de les assignatures precedents d'Equacions Diferencials.

També es pot observar que s'ha produït el mateix fenomen succeït durant la predicció de l'assignatura d'Informàtica, on ambdós models han respòs amb un 100% d'especificitat i un 0% de sensibilitat per l'experiment d'una única variable (n=1). Aquest fet indica que per aquest cas el model tan sols retorna un aprovat, sigui quin sigui el valor de les variables independents.

Igual que en les demés assignatures, es procedeix a afer un anàlisi dels valors de la variable dependent. A partir de la *Figura 14* adjunta es pot mostrar el desnivell d'aprovats i suspesos de la primera convocatòria realitzada de l'assignatura, on han suspès un total de 445 estudiants respecte el total de 2.282, representant el 19,5% de la població d'estudi.

En referència als models de predicció, s'observa una gran similitud en els valors obtinguts per ambdós casos, mostrant resultats pràcticament iguals en la precisió, la sensibilitat i l'especificitat. No obstant, l'exactitud de suspesos en el model de regressió logística és significativament superior respecte l'arbre de decisions, la qual cosa farà decantar la balança alhora d'escollir quin model s'utilitza. En aquest cas es pot observar que per n=6 variables independents el model presenta un màxim relatiu d'aquest indicador, per tant, serà el nombre de variables establert per fer ús del model.

D'aquesta manera, la regressió logística utilitzada amb sis variables significatives encertaria a prop del 82% dels casos predits com aprovat, i aproximadament un 58% de les prediccions de suspesos. Tot això obtenint uns valors propers al 98% d'especificitat i 10% de sensibilitat. Així doncs s'han obtingut uns models molt similars als de la predicció de les assignatures d'Informàtica i Mètodes Numèrics, on s'obté un gran rendiment de pronòstic d'aprovats però es descompensa amb la falta de predicció de suspesos.

Regressió logística de l'assignatura d'Equacions Diferencials												
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	458,73	453,17	450,50	449,23	450,13	449,70	448,30	448,23	448,57	448,23	...	446,97
FP	0,00	5,57	8,23	9,50	8,60	9,03	10,43	10,50	10,17	10,50	...	11,77
FN	112,27	107,13	104,17	103,60	103,43	100,83	101,03	101,33	99,80	99,83	...	97,97
TP	0,00	5,13	8,10	8,67	8,83	11,43	11,23	10,93	12,47	12,43	...	14,30
Precisió	0,8034	0,8026	0,8032	0,8019	0,8038	0,8076	0,8048	0,8041	0,8074	0,8068	...	0,8078
Exactitud (0)	0,8034	0,8088	0,8122	0,8126	0,8132	0,8169	0,8161	0,8157	0,8181	0,8179	...	0,8203
Exactitud (1)	0,0000	0,4885	0,5071	0,4943	0,5201	0,5753	0,5301	0,5165	0,5619	0,5497	...	0,5559
Especificitat (0)	1,0000	0,9879	0,9821	0,9793	0,9813	0,9803	0,9773	0,9771	0,9779	0,9771	...	0,9744
Sensibilitat (1)	0,0000	0,0457	0,0723	0,0774	0,0790	0,1020	0,1003	0,0975	0,1114	0,1112	...	0,1278
AUC	0,6666	0,7037	0,7178	0,7302	0,7333	0,7355	0,7367	0,7417	0,7513	0,7519	...	0,7525
Variable entra	Calc1 nota	Termo conv	Meca1 conv	Quim2 nota	Algeb nota	Calc1 conv	Info1 conv	Termo nota	Calc2 nota	Info1 nota	...	Expre conv

Taula 17: Resultats de la regressió logística de l'assignatura d'Equacions Diferencials

Arbre de decisions de l'assignatura d'Equacions Diferencials												
N	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**...**	**20**
TN	458,73	453,37	451,37	449,10	443,90	442,67	443,13	442,17	438,90	438,73	...	436,03
FP	0,00	5,37	7,37	9,63	14,83	16,07	15,60	16,57	19,83	20,00	...	22,70
FN	112,27	107,63	107,27	106,20	103,73	100,63	101,07	100,03	97,50	97,70	...	95,30
TP	0,00	4,63	5,00	6,07	8,53	11,63	11,20	12,23	14,77	14,57	...	16,97
Precisió	0,8034	0,8021	0,7992	0,7971	0,7924	0,7956	0,7957	0,7958	0,7945	0,7939	...	0,7933
Exactitud (0)	0,8034	0,8082	0,8081	0,8089	0,8107	0,8149	0,8144	0,8156	0,8184	0,818	...	0,8208
Exactitud (1)	0,0000	0,2893	0,3193	0,3154	0,3540	0,4386	0,4365	0,4234	0,4325	0,4246	...	0,4253
Especificitat (0)	1,0000	0,9883	0,9840	0,9791	0,9677	0,9650	0,9661	0,9639	0,9568	0,9565	...	0,9506
Sensibilitat (1)	0,0000	0,0412	0,0445	0,0538	0,0759	0,1040	0,1003	0,1091	0,1314	0,1297	...	0,1509
AUC	0,655	0,6887	0,6943	0,6952	0,6842	0,6918	0,6874	0,683	0,6916	0,6884	...	0,6881
Variable entra	Calc1 nota	Termo conv	Meca1 conv	Quim2 nota	Algeb nota	Calc1 conv	Info1 conv	Termo nota	Calc2 nota	Info1 nota	...	Expre conv

Taula 18: Resultats de l'arbre de decisions de l'assignatura d'Equacions Diferencials

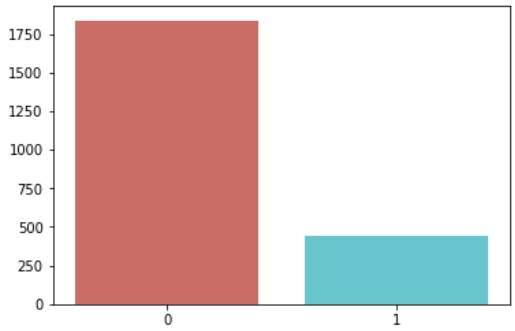


Figura 14: Diagrama de barres dels aprovats (0) i suspesos (1) de la variable Equacions Diferencials

5.4. Variables afegides per la millora dels models

Després d'haver vist i comentat els models de predicció de variables de cada assignatura del tercer quadrimestre, es procedeix a fer un resum de les variables seleccionades per cada model. A la *Taula 19* es mostra la quantitat de variables necessàries per obtenir l'estabilitat en els models de predicció de cada assignatura, així com la descripció d'aquestes i el tipus de model de predicció amb major rendiment. Cal recordar que es busca el model amb el mínim nombre de variables per poder obtenir una major eficiència, ja que d'aquesta manera les poques variables seleccionades aportaran una informació de major qualitat respecte les demés.

Arribats a aquest punt, la metodologia CRISP-DM valora la idea de repetir fases anteriors del procediment per tal de millorar la resposta dels models de predicció. Aplicat al projecte en qüestió, es planteja tornar a comprendre i utilitzar noves variables que es creguin adients per intentar afectar positivament als resultats dels models de predicció.

5.4.1. Noves variables definides

Com ja s'ha comentat a l'inici de la part experimental, les variables utilitzades fins el moment han estat totes corresponents a la les dades de la fase inicial, així que encara no s'ha posat en pràctica cap variable procedent de les dades personals dels estudiants. Per altra banda, al seleccionar la nota més alta de la última convocatòria i el nombre de convocatòries realitzat de cada assignatura, hi ha un seguit de dades acadèmiques que no s'han tingut en compte, com ara la nota obtinguda en els suspesos de cada assignatura.

Assignatura	n estabilitat	Variables										Model
Mètodes Num.	1	Meca1 conv										Regressió logística
Informàtica	4	Info1 nota	Info1 conv	Termo conv	Meca1 conv							Arbre de decisions
Electromagnetisme	10	Termo nota	Meca1 conv	Quim2 nota	Calc2 conv	Termo conv	Meca1 nota	Geom conv	Calc1 nota	Geom nota	Calc1 conv	Regressió logística
Mecànica	4	Termo nota	Algeb nota	Calc1 nota	Meca1 nota							Regressió logística
Materials	7	Quim2 nota	Quim1 nota	Termo conv	Meca1 conv	Termo nota	Expre nota	Meca1 nota				Regressió logística
Eq. Diferencials	6	Calc1 nota	Termo conv	Meca1 conv	Quim2 nota	Algeb nota	Calc1 conv					Regressió logística

Taula 19: Resum de selecció de variables pels models utilitzats a cada assignatura

També s'ha observat que totes les variables fan referència a una assignatura en concret, sense englobar un conjunt de rendiment més continu de l'estudiant. Per aquests motius esmentats s'ha procedit a l'ús i creació de les següents les noves variables definides a continuació.

Nota d'accés: Variable ja existent en les dades personals de cada estudiant, corresponent a la nota obtinguda a les PAU. Es vol considerar ja que proveeix informació prèvia a la primera matrícula de l'estudiant i és fàcil de quantificar degut a que es tracta d'un valor continu. Aquesta també inclou el rendiment del batxillerat o estudis previs, ja que la nota de les PAU té en compte la seva mitjana acadèmica.

Rendiment de primer curs: Nova variable creada, basada en el percentatge d'assignatures no repetides durant el primer curs. Es tracta d'una variable obtinguda a partir de les convocatòries de cada assignatura, per tant és linealment dependent. Per exemple, un estudiant que ha aprovat totes les assignatures a la primera convocatòria obtindrà un rendiment del 100%, mentre que un estudiant que ha cursat totes les assignatures dues vegades n'obtindrà un del 50%.

$$Rendiment\ primer\ curs = \frac{100 \cdot n}{\sum_{i=0}^{n} conv_i}$$

Equació 1: Variable Rendiment del primer curs, on "i" pertany a [0,9] i n al nombre total d'assignatures del primer curs

	Mèt. Num.	Informàtica	Electromagn.	Mecànica	Materials	Eq. Difer.
Nota d'accés	19	21	19	19	18	19
Rendiment 1r	1	1	1	1	1	1
Nota mitjana	1	1	1	1	1	1

Taula 20: Posició de les noves variables en la llista de rellevància de variables independents, segons el model de selecció

NOTA MITJANA: Variable corresponent a la mitjana aritmètica de totes les assignatures cursades durant la fase inicial. Aquesta dada s'obté sumant totes les notes de cada estudiant i dividint-les entre el nombre de convocatòries realitzades.

$$Nota\,mitjana = \frac{\sum_{i=0}^{n} conv_i}{n}$$

Equació 2: Variable Nota mitjana, on "i" pertany a [0,9] i n al nombre total d'assignatures del primer curs

Les noves variables definides s'han introduït al sistema per veure el seu efecte en els models de predicció. En un primer instant, s'ha executat el model de selecció automàtica de variables en cadascuna de les sis assignatures del tercer quadrimestre. Aquest procés s'ha realitzat introduint cada nova variable per separat, per veure com respon el model de selecció amb la nova variable. Per poder analitzar els resultats de manera visual s'ha realitzat la *Taula 20*, que mostra la posició de rellevància de les noves variables segons cada model de selecció.

La taula mostra un fet molt interessant pel projecte, doncs dues de les noves variables definides apareixen com a variable més significativa a tots i cadascun dels models de predicció. Per altra banda, es demostra que la nova variable de nota d'accés no és significativa en cap dels casos, ja que ocupa les últimes posicions en la llista de variables independents rellevants segons el model de selecció. Així doncs, es descarta el seu ús de en el segon anàlisi dels models.

	Rendiment 1r	Nota mitjana	Mètodes Numèrics	Informàtica	Electromag.	Mecànica	Materials	Equacions Diferencials
Rendiment de primer	1	0,8204	-0,2727	-0,333	-0,3699	-0,3492	-0,3018	-0,2870
Nota mitjana		1	-0,2834	-0,3568	-0,4177	-0,4300	-0,3628	-0,3254

Taula 21: Coeficients de correlació de Pearson entre les noves variables i les variables dependents

Abans d'introduir les dues noves variables als models i veure com responen, s'ha procedit a fer un estudi dels seus coeficients de correlació. L'objectiu és poder veure l'impacte de les noves variables independents a la resposta de cada model, així com detectar possibles correlacions entre variables independents. Per poder-ho comprovar, es mostren tots els coeficients correlació de *Pearson* obtinguts a la *Taula 21.*

Hi ha diverses coses a comentar de la taula obtinguda. En primer lloc, s'observa quelcom esperat segons la definició de les dues noves variables, que tinguin una correlació negativa amb totes les assignatures. Aquest fet indica que tant la variable rendiment com la nota mitjana són inversament proporcionals a les variables dependents, ja que en aquest cas el suspès té un valor positiu (1). En segon lloc, s'observa que la variable de nota mitjana obté uns valors més alts en valor absolut respecte la variable de rendiment en tots els casos. En tercer lloc, ambdues noves variables comparteixen un índex de correlació entre elles significativament alt, cosa que pot ser perjudicial per la resposta del sistema, ja que pot generar soroll al resultats obtinguts. Cal recordar que l'objectiu és minimitzar la correlació entre variables independents i maximitzar el valor absolut de la correlació amb les dependents.

Aquestes observacions condueixen a reflexionar sobre utilitzar tan sols una de les dues noves variables definides, ja que l'altra no aportarà un increment significatiu d'informació per millorar la resposta. Donat que el model selecció automàtica de variables atorga la mateixa posició a les dues noves variables en tots els casos d'estudi, es valora la opció d'escollir aquella amb el valor absolut de correlació més alt, en aquest cas la variable de nota mitjana.

5.4.2. Resultats finals

Finalment s'han executat els sis parells de models de predicció amb la nova variable introduïda (nota mitjana) i s'ha vist com responia cadascun d'ells. Donat que ja s'han comentat els trets més característics de cada model als anàlisis previs de cada assignatura, es procedeix a adjuntar la *Taula 21* amb la informació total resumida. Aquesta conté la informació corresponent a l'experiment amb millors resultats. En altres paraules, augmentant el valor de n variables escollides, s'escull la informació de l'experiment que obté uns millors indicadors percentuals en comparació a la resta.

La taula presenta de manera resumida els trets més importants dels resultats dels models de predicció: quin model s'ha utilitzat, quantes variables necessita, quines són aquestes variables i un desglossament dels indicadors percentuals de rendiment. En línies generals es pot observar que el nombre de variables necessàries per obtenir el punt òptim es veu reduït en comparació al resum d'anàlisi de la *Taula 22*.

Pel que fa als indicadors percentuals, si bé és cert que no hi ha una variància significativa respecte l'anàlisi dels primers resultats, almenys aquests s'obtenen amb valors de n inferiors. Per exemple, l'assignatura de Mètodes Numèrics ha obtingut una sensibilitat del 15% amb tan sols una variable, mentre que als resultats anteriors no es superava el 8% en cap dels vint experiment. El mateix passa amb altres assignatures com Electromagnetisme o Equacions diferencials, on s'han necessitat tan sols dues variables per poder obtenir el punt òptim dels models, quan abans se'n necessitaven deu i set respectivament.

Per acabar, de tots els resultats obtinguts es considera que els models de predicció per les assignatures de Mecànica i Electromagnetisme són els que han obtingut millors resultats, ja que presenten un gran equilibri en el conjunt d'indicadors percentuals. Una bona prova d'aquest fet és el seu alt paràmetre AUC obtingut, que mostra una bona relació entre la predicció d'aprovats i suspesos d'ambdues assignatures. L'assignatura d'Informàtica també presenta un alt valor AUC, tot i que la sensibilitat obtinguda no és gaire elevada. Un bon model de predicció es caracteritza per poder predir correctament el conjunt sencer de solucions possibles, tant els valors positius com els negatius.

El desglossament dels resultats obtinguts en els models finals es pot veure amb detall a l'annex adjunt al treball.

Assignatura	Mètodes Numèrics	Informàtica	Electromag.	Mecànica	Materials	Eq. Diferen.
Model predicció	Arbre D.	Regressió	Regressió L.	Regressió L.	Regressió L.	Arbre D.
n d'estabilitat	1	4	2	5	3	2
TN	489,2	438,2	324	173,4	344,4	443,2
FP	8,6	15,5	53,17	103,9	47,03	15,5
FN	62,1	93,5	105,3	73,5	123,3	97,1
TP	11,1	23,8	88,5	220,2	56,3	15,2
Precisió	0,8762	0,8092	0,7224	0,6893	0,7017	0,8028
Exactitud (0)	0,8873	0,8243	0,7548	0,7026	0,7365	0,8204
Exactitud (1)	0,5473	0,6072	0,6251	0,6795	0,5462	0,4815
Especificitat (0)	0,9828	0,9659	0,8591	0,6256	0,88	0,9661
Sensibilitat (1)	0,1519	0,2037	0,4572	0,7499	0,3138	0,1344
AUC	0,7266	0,7647	0,7601	0,7604	0,7287	0,7133
Variables	Nota mitjana	Nota mitjana	Nota mitjana	Nota mitjana	Nota mitjana	Nota mitjana
		Nota Info	Termo Nota	Nota Termo	Nota Quím. 2	Nota Càlcul 1
		Conv Info		Nota Algebra	Nota Quím. 1	
		Conv Termo		Nota Càlcul 2		
				Nota Mec. Fon.		

Taula 22: Resum dels resultats finals obtinguts amb la nova variable, segons l'experiment que estabilitza la resposta

Per acabar, cal esmentar que s'ha procedit a realitzar un últim procediment per la comprovació de resultats. Per tal de corroborar les suposicions fetes alhora de seleccionar les noves variables, s'ha decidit executar els models de predicció tant amb cada una d'elles per separat com amb les tres introduïdes a la font de dades independents.

En aquest cas s'ha pogut confirmar que la millor resposta obtinguda ha estat la que incloïa tan sols la nota mitjana. La variable rendiment de primer aporta una informació tan similar que no afecta als resultats, reduint la eficiència del model degut a que s'utilitza un major nombre de variables. Com ja s'ha comentat, la variable nota d'accés no apareix mai entre les variables significatives. Aquest fet destaca la importància de complementar el mètode de selecció automàtica de variables amb altres mitjans matemàtics i estadístics per aprofundir la informació sobre les variables d'estudi.

6. Planificació i pressupost

Un aspecte important que ha d'incloure tot projecte és l'estudi econòmic i la planificació d'aquest. Això permet que qualsevol persona, empresa o entitat interessada en la realització del projecte pugui disposar de la informació essencial sobre la inversió que caldria efectuar i el termini de temps necessari per la seva realització.

Començant per la planificació, la *Figura 15* mostra el diagrama de *Gantt* de les diferents tasques dutes a terme durant aquest projecte. Aquest esquema s'ha intentat seguir de manera rigorosa per tal de dedicar el temps just a cada tasca i poder finalitzar el treball en el termini establert. Es pot veure que la durada total del projecte, des de l'assignació fins la seva presentació, ha estat de cinc mesos.

Pel que fa al pressupost del projecte, s'ha procedit a fer un estudi de tots els possibles costos generats, desglossats a la *Taula 23* per comptar amb un suport visual. Donat que el treball realitzat forma part de les ciències de la computació, aquest es pot exercir tan sols amb suport informàtic, sense la necessitat manufacturar cap tipus de producte ni realitzar cap experiment físic. Això implica que un elevat percentatge del pressupost total provingui de les despeses de personal, les quals inclouen les diferents professions descrites a continuació.

- **Programador/s:** Personal especialitzat en el tractament de dades per projectes *Data Mining*, el qual ha de ser capaç de preparar i transformar les dades proveïdes, construir els diferents models de predicció i selecció exposats i desar la resposta obtinguda amb un format llegible. També s'exerciran aquelles modificacions proposades pels analistes.

- **Analista/es:** Es requereix personal analista capaç d'interpretar els resultats obtinguts per extreure'n conclusions fiables, així com identificar els diferents factors que afecten a la resposta del sistema. També haurà de proposar possibles solucions per la millora del projecte, com ara la creació de noves variables. Aquesta informació serà transmesa de nou al/s programador/s per dur a terme la seva implementació.

Fases	Activitats	Febrer 2019	Març 2019	Abril 2019	Maig 2019	Juny 2019	Juliol 2019
Inici del projecte	Nominació del projecte						
	Definició i comprensió del problema						
	Recerca d'informació						
	Instal·lació del programari						
Comprensió i preparació de dades	Introducció de dades al programari						
	Comprensió de dades i definició d'estructura a obtenir						
	Transformació de dades						
Modelatge	Estudi de models de predicció						
	Estudi de models de selecció de variables						
	Creació de models desitjats						
Anàlisi de resultats	Validació dels models creats						
	Definició i introducció de noves variables						
	Validació de resultats finals						
Confecció de la memòria i presentació	Recerca d'informació complementària						
	Redacció de la memòria						
	Elaboració i resum de taules de resultats						
	Presentació del projecte						

Figura 15: Diagrama de Gantt del projecte

TAULA DE PRESSUPOST TOTAL DEL PROJECTE			
DESPESES DE PERSONAL			
	Dedicació (hores)	**Preu (€/hora)**	**Cost (€)**
Programació	160	20	3.200
Anàlisi	120	20	2.400
Gestió de projectes	90	35	3.150
SUBTOTAL	370	-	8.750
DESPESES MATERIALS			
	Dedicació (mesos)	**Preu (€/any)**	**Cost (€)**
Ordinadors de treball	5	1.200	500
Llicències de programa	12 (mínim d'un any)	99	99
Despeses d'oficina	5	1.520	633,33
SUBTOTAL	-	-	1.232,33
TOTAL	-	-	**9.982,33**

Taula 23: Pressupost econòmic del projecte

- **Enginyer/s de projectes:** Per dur a terme un projecte d'enginyeria com aquest es necessita la figura d'un *Project Manager*, el qual determini la direcció i objectius del projecte i en faci un seguiment constant. Aquest serà responsable d'interpretar les accions que s'han de dur a terme i coordinar l'equip d'analistes i programadors.

Per altra banda, les despeses materials també representen una part a tenir en compte en el pressupost econòmic del projecte. Cal destacar que el treball té una planificació total de cinc mesos, per tant, les despeses adjuntes es consideren en relació al lloguer de material durant aquest període de temps, no a la seva adquisició permanent.

- **Material informàtic:** Com bé s'ha comentat, el projecte forma part de les ciències de la computació, per la qual cosa es necessiten ordinadors amb suficient memòria i capacitat per poder processar les dades amb diferents programes informàtics. Aquest cost inclou les despeses per amortització del material.

- **Llicències:** De cada programa informàtic emprat s'ha de comptabilitzar el cost provinent de les llicències. Donat que el projecte es realitza mitjançant el llenguatge d'accés gratuït *Python*, tan sols s'ha de tenir en compte el cost del paquet *Office*, utilitzant el programa *Word* per la redacció de l'informe i *l'Excel* per l'estructura de resultats.

- **Despeses d'oficina:** Les despeses d'oficina fan referència al lloguer del local o espai on es durà a terme el projecte, tenint en compte els costos associats a la llum, serveis i el material d'oficina (cadira ergonòmica, impressora, paper, etc.).

Tot plegat suma un pressupost total de 9.982,33 € bruts, dels qual 8.750€ pertanyen a les despeses de personal i 1.232,33€ a les despeses materials i d'oficina.

7. Impacte mediambiental

Un últim aspecte a tenir en compte en la realització d'un projecte és l'impacte generat al medi que l'envolta. En aquest cas, es torna a fer èmfasi en el fet de que el treball en qüestió tan sols s'executa una transformació d'informació generada mitjançant material informàtic. Així doncs, l'impacte ambiental generat és pràcticament nul en comparació a qualsevol altre projecte, ja que en aquest no es realitza cap acció pràctica més enllà del treball d'oficina. D'aquesta manera tan sols es té en compte l'impacte generat a nivell de recursos informàtics i despeses energètiques.

Per una banda, segons dades de la *Comissió Europea*, cada ordinador emet entre 52 i 234 grams equivalents de CO_2 per hora d'ús, considerant una potència d'entre 80 i 360 watts. Tenint en compte un ús total d'ordinador de 370 hores, això suposa enviar una mitjana de 52,9 kg equivalents de CO_2 en la realització d'aquest projecte. **[9]**

Addicionalment s'ha de tenir en compte el consum energètic provinent de les llums que il·luminen l'espai de treball. Considerant de nou les hores de treball definides, una taxa d'ús del 50% (degut a l'aprofitament de la llum solar) i un total de tres bombetes de baix consum, aquestes emetran fins a 3,9 kg de CO_2. Aquest valor s'ha obtingut basant-se de nou en la taula de valors de la *Comissió Europea*, que estableix una emissió de 7 g/hora de CO_2 per les bombetes de baix consum.

Tot i que és un impacte ambiental a tenir en compte, no suposa un valor significatiu en comparació a qualsevol altre tasca laboral, ja que no es genera cap altre residu més enllà del consum elèctric. El material emprat també pot ser reutilitzat per altres tasques un cop finalitzi el projecte.

8. Conclusions

8.1.1. Objectius assolits

Per acabar, es pot dir que s'han complit els objectius principals definits a l'inici del treball. S'han creat i analitzat diversos models de predicció del rendiment acadèmic dels estudiants d'enginyeria industrial de l'ETSEIB. Per fer-ho s'ha adaptat la metodologia CRISP-DM a les dimensions del projecte, una de les tècniques més utilitzades en l'àmbit de la mineria de dades.

A nivell tècnic, s'han explorat i reproduït diferents eines de mineria de dades utilitzades en la predicció classificatòria, tals com la regressió logística o els arbres de decisions. Tanmateix s'ha comprovat l'eficiència d'un mètode de selecció automàtica de variables per filtratge, el qual ha ajudat a obtenir una millora de resultats de predicció de l'aprovat o suspès de les assignatures del tercer quadrimestre. Tot i que tan sols s'ha pogut executar un mètode de selecció, aquest ha estat aplicat a diversos models de predicció de variables, comptant també amb altres recursos i eines per estudiar una metodologia de selecció de variables.

Pel que fa al desenvolupament del projecte, s'han pogut crear diferents models de predicció que es podrien utilitzar per aquell qui ho desitgi o necessiti. Analitzant la viabilitat d'aquests models s'ha pogut identificar els avantatges i inconvenients a tenir en compte de cadascun, així com proposar possibles millores de futur.

De tots els resultats obtinguts n'hi ha hagut alguns de prou eficients, els quals demostren poder utilitzar els models en casos reals per la predicció de l'aprovat o suspès d'algunes assignatures com Mecànica, Electromagnetisme o Informàtica. També s'han pogut identificar aquelles assignatures que presentaran més dificultats alhora de predir resultats.

Cal tenir en compte les limitacions matemàtiques dels models, doncs hi ha una infinitat de factors humans que no es poden tenir en compte ni es poden enregistrar en forma de dades.

8.1.2. Nivell personal

A nivell personal es valora molt positivament la realització del treball, doncs s'han assolit alguns objectius addicionals proposats a l'inici del projecte. S'ha pogut obtenir una bona base de coneixement per la realització de projectes *Data Mining*, així com entendre tot el procés de transformació de dades pel seu futur anàlisi.

També cal dir que ha estat tot un plaer tornar a utilitzar el llenguatge de programació *Python*, el qual permet arribar a metes molt llunyanes partint de la senzillesa i fàcil comprensió del programa. Aprofitant aquest tema, cal esmentar que una gran part del projecte ha consistit en la creació del codi del programa, el qual suposa un gran esforç de comprensió, aprenentatge i implementació que no es veu plasmat en la redacció de la memòria.

Es valora molt l'increment de coneixement adquirit en el llenguatge de programació, concretament en l'ús de les llibreries *pandas* i *scikit-learn*. Tot plegat permet poder dur a terme un anàlisi de dades més ordenat, hàbil i fiable en un futur professional en el sector de l'enginyeria.

9. Futur del projecte

9.1.1. Fase d'implementació

La fase final del procediment CRISP-DM per projectes de mineria de dades estableix posar sobre la taula els coneixements adquirits per tal de poder prendre decisions per resoldre un problema. També es recomana crear un informe on es recopili tota la informació de l'estudi realitzat, documentant els resultats de manera comprensible per tal de poder ser usats per un futur usuari en cas necessari.

Pel que fa a l'increment de coneixements, la realització d'aquest projecte ha permès crear un programa informàtic que disposa de diferents tipus de models de predicció per l'aprovat i suspès de les assignatures del tercer quadrimestre de l'ETSEIB. Per aquest simple fet, qualsevol estudiant o individu que vulgui utilitzar aquests models tan sols haurà d'introduir les variables requerides i executar el programa.

No obstant, s'haurà de tenir en compte tots els aspectes comentats durant la realització del projecte. La eficiència dels models varia en funció de l'assignatura a predir, tenint una major precisió per la predicció d'aprovats respecte la de suspesos. El millor model obtingut ha estat el de la predicció de l'aprovat o suspès de l'assignatura de Mecànica, on aproximadament 7 de cada 10 resultats són encertats correctament, tant per aprovats com per suspesos. Aquest fet es podria aprofitar per dur a terme els models de predicció en casos reals, i prendre posteriors conclusions sobre el rendiment acadèmic dels estudiants i el sistema educatiu de l'escola.

En referència a l'informe a redactar, aquesta pròpia memòria de treball de fi d'estudis pot servir per futurs usuaris. S'han definit les eines usades i s'ha mostrat en tot moment els valors dels resultats obtinguts, representant en diverses taules els valors que s'han cregut més rellevants. També s'han redactat els passos experimentals duts a terme i s'han adjuntat uns annexos on s'explica al detall el programa codificat.

9.1.2. Propostes de millora

Com tot treball de fi d'estudis, aquest projecte té alguns límits d'abast degut a diferents

factors, ja siguin per falta de temps, de coneixements o de recursos tecnològics.

En primer lloc, els propers passos a seguir serien utilitzar diferents metodologies de selecció de variables, com ara el mètode *wrapper*. Aquesta eina permet generar múltiples combinacions de subconjunts de variables independents i analitzar la seva interacció amb les variables dependents. D'aquesta manera es fa una selecció de variables més fiable, ja que el mètode de filtratge utilitzat tan sols té en compte la relació entre cada variable independent particular i la dependent a predir. Això pot comportar passar per alt possibles interaccions entre variables d'entrada que afectin negativament la resposta.

Una altra proposta de millora pel futur del projecte seria fer una recol·lecció de noves dades dels estudiants. Posant sobre la taula l'exemple redactat anteriorment sobre l'estudi de la taxa d'abandonament universitari a l'Argentina, es podria ampliar la informació dels estudiants amb variables com el nivell acadèmic adquirit pels pares o l'edat de l'estudiant. També cal dir que en un moment es va plantejar la idea d'utilitzar els codis postals de les dades personals per veure si la localització del domicili de l'estudiant podia influir en el rendiment acadèmic. Donat l'extens rang de valors que pot adquirir aquesta variable es va decidir no utilitzar-la, tot i que també es podria aprofundir per aquesta direcció.

Aquestes són les propostes que es creuen més factibles per l'ampliació de l'estudi amb l'objectiu de millorar els resultats obtinguts. Altres noves propostes sempre poden ser benvingudes, com ara afegir nous models de predicció o inclòs utilitzar una altra metodologia pel procediment de mineria de dades. De totes maneres, s'ha de tenir en compte la línia del projecte per tal de poder millorar-lo, podent comparar així les noves propostes amb els resultats obtinguts fins al moment. Si les propostes s'allunyen massa de la trajectòria del treball es podria plantejar la opció de realitzar un nou projecte independent.

BIBILIOGRAFIA

Referències bibliogràfiques

[1] D.L. OLSON, D. URSUN: *Advanced Data Mining Techniques.* Spring 2008, Berlin

[2] L. VALÍA, J. ROSTAGNO, J. M. MOINE, C. BIGATTI, F. M RIVA, E. AMAR: *Modelo de diserción universitaria en la Universidad Tecnológica Nacional Facultad Regional Rosario*. Año 2017. [http://sedici.unlp.edu.ar/handle/10915/61720]

[3] J.A. GALLARDO: *Metodología para la definición de requisitos en proyectos de Data Mining*. Año 2009 [http://oa.upm.es/1946/]

[4] A. KUMAR: *Learning predictive analytics with Python*. February 2016 [https://books.google.es/books?id=Ia5KDAAAQBAJ&hl=es]

[5] A. NAVLANI: *Understanding Logistic Regression in Python*. September 2018 [https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python]

[6] R. SAXENA: *Building decisión tree algoriothm in Python with SciKit Learn*. February 2017 [http://dataaspirant.com/2017/02/01/decision-tree-algorithm-python-with-scikit-learn/]

[7] R. SHAIKH: *Feature selection techniques in Machine Learning with Python*. October 2018 [https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e]

[8] P.R. DE LOS SANTOS: *Machine Learning a tu alcance: La matriz de confusión*. Enero 2018 [https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/]

[9] Entitat Ecoembes: *Los ordenadores también emiten CO2.* Abril 2016 [https://www.ecoembes.com/es/planeta-recicla/blog/los-ordenadores-tambien-emiten-co2]

Bibliografia complementària

Font d'informació KdNuggets: [https://www.kdnuggets.com]

Portal de l'usuari de Pandas: [https://pandas.pydata.org/pandas-docs/stable/]

Portal de l'usuari SciKit-learn: [https://scikit-learn.org/stable/]

Portal de l'usuari Seaborn: [https://seaborn.pydata.org/]

Informació de la llibreria Numpy: [https://docs.scipy.org/doc/numpy/user/quickstart.html]